Master's thesis

presented to the Faculty of Arts and Social Sciences

of the University of Zurich

for the degree of

Master of Arts UZH in Social Sciences

Underestimating the Instruction

A Meta-analysis on Item Count Technique and Crosswise Model

Author: Antonia Velicu

Student ID Nr.: 17-725-953

Examiner: Prof. Dr. Heiko Rauhut

Supervisor: Julia Jerke

Institute of Sociology

Submission date: September 30, 2019

Abstract

Survey respondents tend to present themselves in a more favorable light, especially when being asked unpleasant questions. This so-called social desirability bias introduced by sensitive questions often distorts survey responses. As a remedy research draws on indirect questioning formats that aim to protect respondents' privacy and ensure their anonymity. Two prominent examples of such techniques are the Crosswise Model (CM) and the Item Count Technique (ICT). Both methods follow unconventional structures using group answers or known distributions to mask individual answer but that also require long, complex and dense instructions. Previous research has suggested that ICT and CM produce more truthful answers, however they impose a higher cognitive burden on respondents. Although, it is commonly believed that respondents fully understand and follow these more demanding instructions, recent research suggests that this is not always the case. To further investigate this notion, I conduct a meta-analysis of the ICT and CM and analyze the instructions of these methods to answer two core questions: First, how do the implementations of the Item Count Technique and the Crosswise Model differ across studies? Second, how do specific characteristics (i.e., the instruction) of the techniques affect their performance? The meta-analysis indicates mixed results on the performance of the techniques. The CM tends to perform better than the ICT. ICT works best when asked in face-to-face interviews, the sensitive item phrased as a socially undesirable one, and the non-sensitive items chosen from the same contextual background. ICT instructions with too many words and not many word repetitions appear to have a negative influence on its outcome. The results of this research have implications for researchers and practitioners working with these techniques, but also for the broader field measuring and analyzing sensitive characteristics in surveys.

Keywords: Item Count Technique; Crosswise Model; Meta-Analysis; Cognitive Burden; Survey Methodology; Instructions

Acknowledgments

I would like to thank Heiko Rauhut and Julia Jerke for giving me the possibility to write my master thesis and for supervising it, and David Johann for his helpful comments during the process. I'm also very grateful for the inspiring discussions and support from Justus Rathmann, Isabel Raabe and Alexander Ehlert.

I want to thank Roxana & Eugen for their unconditional love, and moral and financial support, Berni for everything he has done for me without ever questioning, and last but not least my friends whom I owe more than just my mental health.

Contents

List of Figures

List of Tables

1 Introduction

Do people believe Obama is a Muslim? Do researcher fabricate data? Are Persian students promiscuous? Does the Chinese population have confidence in their national government? Did people vote during the last election? All these questions are taken from conducted surveys but can researchers expect to get truthful answers? It can be assumed, that these questions generate response errors, resulting in having a negative impact on the data quality (Krumpal (2013)).

The aforementioned questions address sensitive topics. With the advent of modern surveys, the interest to investigate sensitive topics also rose, but modern surveys face multiple problems: Sampling problems can occur as a result of hard-to-contact or unlisted groups of people; non-response problems arise due to the harmfulness of the topic; or problems with the quality of answers are caused by social desirability and memory issues (Lensvelt-Mulders (2008): 477f). The final complication pertains to the case, where the answers of respondents are in the sense biased, to met alleged expectations of researchers toward themselves (Bogner and Landrock (2015)).

Questions are considered as sensitive if answering truthfully can have considerable consequences (Tourangeau and Smith (1996): 276). From a theoretical point of view, the three aspects of sensitivity are (Tourangeau and Yan (2007): 860): (i) intrusive, (ii) threat of disclosure and (iii) social desirability. The first aspect pertains to the case, when a question per se is too private or offensive and thus *intrusive*. Further, a casual question can turn into a sensitive one, if the threat of a third party finding out the answer is given (*threat of disclosure*). For example, asking a teenagers about smoking behavior in a setting with friends or asking them in a setting where the parents can find out easily, will probably distort the answer. The third aspect of sensitivity involves the nature of the answer. The concept assumes there are clear social norms what behavior or attitude is acceptable and what is in deviation of the norms and thus *socially undesirable* (iii). For example, a question about voting is per se not sensitive, but there are norms, that citizens should go voting because of their civic duty. For a person who went voting, a question about their voting behavior is not sensitive. For a person who decided against carrying out their civic obligation, the question turns into a sensitive one as a truthful answer would be socially undesirable (ibid.)

There are many ways to determine what is sensitive, one approach is to look at results of previously conducted surveys (Seibert (2019), Krumpal (2013)). For instance, a parameter could be to identify a higher item non-response, e.g., to look at questions which participants are more often reluctant to answer (Lensvelt-Mulders (2008); Tourangeau and Yan (2007)). Additional Krumpal (2013) suggests researchers can ask themselves which questions can be a threat and in the following treat them as sensitive (ibid: 2027). Another approach is to let experts or laymen rate the sensitivity of the question on a scale or to consult external checks if available (f.e. cross checking with actual records, ibid.; Bradburn, Sudman, and Wansink (2004); Coutts and Jann (2011)).

In order to avoid potentially stigmatizing situations, researchers developed multiple strategies. On a methodological level, small changes to the survey setup help with minimizing the dilemma: e.g., different framing, adjustment in the survey format, or different ordering of questions (first non-threatening, then sensitive and potentially incriminating and then friendly questions) along with an adjusted content and wording of the question such as the usage of forgiving wording (Lensvelt-Mulders (2008): 468f). Furthermore, the use of self-administered questionnaires instead of face-to-face interviews has proven to be a valuable method in order to reduce reporting errors (Yan and Cantor (2019): 49f). Another option is the usage of indirect questions, which might convince with additional privacy protection and no forced disclosure of individuals (ibid.; Bradburn et al. (2004)). Thus, these techniques promise to overcome the limitations of direct questions and reduce over- or underreporting of participants.

This thesis focuses on two prominent indirect question techniques the Crosswise Model (CM) and the Item Count Technique (ICT). The CM is chosen as it is a revised and new, more promising variant of the original Randomized Response Technique (Warner (1965)); one of the first indirect question techniques. The ICT is chosen because it the most frequently applied technique and there are studies, indicating the superiority of the ICT over all other techniques (f.i., Holbrook and Krosnick (2010); Ahart and Sackett (2004)). This is further emphasized by the vast number of variations, e.g., Item Sum Technique by Trappmann, Krumpal, Kirchner, and Jann (2014) for continuous variables, Person Count Technique by Grant, Moon, and Gleason (2012) for a list of people, Longitudinal Item Count Technique by Gaia and Al Baghal (2019) for longitudinal

results, the Single Sample Count by Petróczi et al. (2011) with non-sensitive items, where the distribution is known and many more.

Both, the Crosswise Model as well as the Item Count Technique come in an rather unconventional package and differ significantly from regular questions in surveys. The CM presents at its participants two questions at the same time. One sensitive and one question unrelated to the controversial topic, e.g., whether the birthday of the respondent falls in the first three months of the year. The respondent can than choose to answer: *«the answers to both questions are the same (no or yes)»* or *«the answers to both questions are different (one yes one no)»* and must at no point admit to a stigmatizing behavior. By knowing the distribution of answers of the unrelated question researchers can estimate responses to the sensitive question (Yu, Tian, and Tang (2008), Jann, Jerke, and Krumpal (2012)). Secondly, the Item Count Technique (ICT) hides the controversial item between multiple non-sensitive statements and the respondents have to indicate the number of items applying to them without disclosing which ones exactly. Participants are split into two groups, a treatment group (full list of items) and a control group (list without the sensitive item). The difference of the average number of items that apply to participants between both groups is a direct indicator of the prevalence of the sensitive item (Droitcour et al. (1991), Tsuchiya, Hirai, and Ono (2007)). To summarize, both techniques use an unconventional structure which makes them more difficult to comprehend. Therefore, ICT and CM often come with extra instructions to participants. And yet, it is unclear how those instructions influence the outcome of the survey.

Contribution and Research Question: The Crosswise Model and the Item Count Technique aim at reducing under- and over-reporting by offering privacy and anonymity to participants. Both techniques have shown mixed results in past surveys leading to a general dispute, whether they are appropriate to ask sensitive questions or not (e.g. Johann, Thomas, Faas, and Fietkau (2016), Hoffmann, de Puiseau, Schmidt, and Musch (2017)). It is thus essential to investigate those techniques in detail and consider their implementations. Therefore, this thesis aims at answering:

> *(RQ1) How do the implementations of the Item Count Technique and the Crosswise Model differ across studies?*

Furthermore, it is not only essential to understand the different implementations of both techniques but also what characteristics could have an effect on their success. Therefore, another object of this thesis is to give insights on the following question:

> *(RQ2) How do specific characteristics (i.e., the instruction) of the techniques affect their performance?*

This questions are answered by conducting an exploratory meta-analysis comparing over 150 different implementations of both techniques over 35 topics in more than 50 countries.

First, Chapter 2 explains the Item Count Technique and Crosswise Model more in depth and shows what differences in applications where chosen by previous research. Then, in Chapter 3, a short state-of-the-art along with the gap in research are presented and afterwards theoretically explained why instructions could have an influence on the performance. I chose the tool of an meta-analysis to investigate the issue. Chapter 4 clarifies the method, demonstrates how the bibliography was complied and which variables were coded. The explorative results are consistent of two parts and are presented in Chapter 5. The first part gives an overview of CM studies and specific characteristics are investigated in a further descriptive analysis. The second part leads with a descriptive analysis of the ICT and sequences with a more in-depth analysis of the method. In particular the influence of the instruction of the method is considered. The results are discussed with the literature in the same chapter. Finally, I conclude with some recommendations for further research.

2 Indirect Question Techniques

A behavior is socially desirable if it is motivated by the desire to gain social approval. Social approval is only given when behaving in a certain culturally accepted way (Marlowe and Crowne (1964): 40). This phenomenon is problematic for survey research as it leads to bias. Social desirability bias manifests in non-response or responses divergent from the truth. It arises due to the attempt to behave compliant to norms, values, roles, or expectations (see Bogner and Landrock (2015)). If a respondent is attempting to meet alleged expectations of the researcher resulting in manipulated answers, their answer is biased because of social desirability. The alleged expectations can be wired in two ways (Krumpal (2013)): The item in question can be (i) socially undesirable (f.e., criminal behavior) or (ii) socially desirable (f.e., blood donation). The bias becomes proportionally bigger as the sensitivity increases (ibid.). The ICT and the CM proposed solutions to the problem of social desirability bias, as they promise more anonymity. Respondents do not have to answer sensitive questions directly and admit a socially (un-) desirable behavior, but instead can answer indirectly. A desired result of indirect question techniques thus are more truthful answers. In this section the two methods subject of this thesis (ICT and CM) are explained and relevant research is discussed.

2.1 Item Count Technique

While most of the studies agree on the fact that Droitcour et al. (1991) was one of the first studies, that applied the Item Count Technique, not all agree on its theoretical originality. Smith and Street (2003) sees it as a variant of the Block Total Response Method (by Raghavarao and Federer (1979)), others locate the origin in Miller's Dissertation (1984). The technique runs by many names, for instance: Unmatched Count Technique (e.g. Nuno and John (2015); Dalton, Wimbush, and Daily (1994); Ahart and Sackett (2004); Coutts and Jann (2011)), Randomized Lists Technique (e.g. Zimmerman and Langer (1995)) or List Experiments (Blair and Imai (2012); Blair, Coppock, and Moor (2018); Ahlquist (2018)). Nevertheless the strategy is the same: Persons never report directly a behavior or attitude of sensitive matter but instead are requested to count and declare the total amount of multiple statements. After an

introduction to the proceedings of the ICT, respondents recieve a list of statements and are asked to count the number of items, that apply to them (hence the name *Item Count*). Among multiple neutrally formulated items, one is about the sensitive topic in question. For instance, if the sensitive topic of interest is plagiarism and the respondent group are students, the following ICT can be asked (inspired by Coutts, Jann, Krumpal, and Näher (2011)):

`Here is a list of five statements that are true for some students, but not for others. Please indicate, how many of these statements are true for you. Please don't write down which ones, but only how many:`

- `On a typical university day, I commute more than 50 km one way`
- `I regularly participate in my research seminars`
- `I have a personal subscription for at least one scientific journal`
- `I predominantly use Mac for my studies`
- `I have deliberately concealed a quotation in my master thesis`

`Please write down below how many of the statements are true for you.`

In order to calculate the prevalence rate of the sensitive item, the sample of respondents is divided into two groups. The first group (treatment group) obtains a long version of the question including all items (e.g. the one appearing in the example). In this case it includes four neutral and one sensitive item (the last item). The second group (control group) receives the same question without the sensitive item (Droitcour et al. (1991)), in this case four non-sensitive items. The next step is to calculate the group means of the treatment and control group answers and subtract them. A difference in means then equals the prevalence rate of the sensitive item. For example, if $\overline{x}_{TG} = 2.45$ and $\overline{x}_{CG} = 1.67$ then $\overline{x}_{TG} - \overline{x}_{CG} = 0.78$. Accordingly to this case, 78% of the respondents have deliberately concealed a quotation in their thesis. Instead of individual answers, researchers acquire information on the group level of the unobserved likelihood of positive answers to the sensitive question (ibid., Tsuchiya et al. (2007)).

2.1.1 Premises of the ICT

The list of non-sensitive items has to be constructed thoughtfully in order to avoid so-called floor or ceiling effects (Holbrook and Krosnick (2010)). If non-sensitive items are very general and unspecific, resulting in all of them applying to all respondents, the anonymity of a person who the sensitive item would apply to can not be guaranteed anymore. This person would have to respond with the maximum (the total sum) of the items and thus directly answer the sensitive question (*ceiling effect*). Something similar applies to the case where non-sensitive items are obviously too constrained. Hiding the sensitive item between items which only apply to a very restricted population increases the hurdle to admit the sensitive statement (*floor effect*). The ideal scenario would be two concealed negative correlated items, one which applies to one part of the population and excludes the other item and vice versa (ibid.; Tsuchiya et al. (2007), Glynn (2013): 163f). For example Glynn (2013) implemented a negative correlation in their list experiment between the items *«Teaching intelligent design along with evolution in public schools»* and *«Making it legal for two men to marry»*, as respondents who dislike the teaching of intelligent design[1] are less likely to dislike gay marriage and vice versa (ibid.: 164).

The ICT is based on two assumptions which should not go without discussion: the no-liars-assumption and the no-design-effect-assumption (Imai (2011); Blair and Imai (2012)). Whether the ICT works or not is crucially dependent on these two assumptions. The no-liars assumption refers to the fact, the technique is only able to show a difference in prevalence if respondents are willing to answer truthfully thus either admit socially undesirable or deny socially desirable behavior. This is a rather strong presumption and fails in many practical experiments (Li (2019)). The no-design-effect assumption refers to the design of the question itself: the mere inclusion of another item (the sensitive item) should have no effect on the response to the short list (Imai (2011); Blair and Imai (2012)). While there are statistical tests to include the no-design-effect-assumption into calculations and hence to detect certain forms of violations, verifying the no-liars-assumption is more problematic. Li (2019) recently found a way to relax the assumption by introducing parameters which capture the proportions of liars.

[1] Intelligent design is a pseudoscientific argument for the existence of God. Followers believe that certain features of the universe and of living things are best explained by an intelligent cause, not an undirected process such as natural selection (Discovery Institute last accessed on 26.09.2019)

2.1.2 Applications of the ICT

The geographical areas in which the ICT is used are many. There are studies from Liberia (Moseson et al. (2015)) to the US (e.g., Rosenfeld, Imai, and Shapiro (2016)) and Korea (Kim and Kim (2016)). The diffusion in terms of content is also rather broad, encompassing topics from unethical behavior in work environment (Dalton et al. (1994)) to risky sexual behaviour and substance abuse (LaBrie and Earleywine (2000)), hate crimes (Rayburn, Earleywine, and Davison (2003)) and blood donation (Tsuchiya et al. (2007)). For an overview see section 5.2.

These are not the only parameters that can be altered, the way the ICT is implicated can also vary across studies. The possibly most obvious alteration is the number of non-sensitive items. Researcher vary the number of non-sensitive Items from three (f.e., Frye, Gehlbach, Marquardt, and Reuter (2017); Comşa and Postelnicu (2012); Holbrook and Krosnick (2010)) up to seven (Roberts and John (2014)). Tsuchiya et al. (2007) recommend the use of less non-sensitive items – precisely three – as they promise enough anonymity, while sticking to a simple calculation. Other researchers find more non-sensitive items better in order to avoid floor and ceiling effects (Sheppard and Earleywine (2013); Dalton et al. (1994); LaBrie and Earleywine (2000); Rayburn et al. (2003)).

What types of sensitive and non-sensitive items are used in the study is another parameter, that researchers tend to vary. They either derive from the same contextual background, or not. Some argue that having chosen all items from the same topic is less unconventional (Frye et al. (2017); Thomas, Johann, Kritzinger, Plescia, and Zeglovits (2017)), may cause less suspicion (Hubbard, Casper, Lessler, et al. (1989)) and is thus advisable. Others disagree and choose non-sensitive items from a different background ((f.e., Rayburn et al. (2003); Holbrook and Krosnick (2010); Glynn (2013); Davis et al. (2019)).

The amount of information given to the respondents throughout surveys can also vary. In earlier years of the ICT, researchers used to do a full briefing with their respondents about the goals of the research and an extensive explanation and examples of the ICT (e.g. Dalton et al. (1994): 821). Later Ahart and Sackett (2004) tested the same items in two groups with the only difference being the inclusion of an explanatory introduction.

The results showed partly significantly higher rates in the sample with the explanation of the ICT (Ahart and Sackett (2004): 107). Fairly soon after this insight Tsuchiya et al. (2007) show in their comparative study similar results. Possible reasons for this outcome are discussed in next chapter (section 3.3).

More recently, researchers found differences concerning the mode of how the ICT was conducted (Rosenfeld et al. (2016)). Their ICT performed differently in a telephone interview as opposed to a questionnaire. The survey mode has proven to make a difference as a strategy to reduce reporting error, in direct and indirect question techniques (Yan and Cantor (2019): 49f). Researchers found that facing a person and having to concede to a stigmatizing attitude or criminal behavior a person is more concerning with respect to privacy than using self-administered surveys (see ibid., Tourangeau and Yan (2007); Preisendörfer and Wolter (2014); Tourangeau and Smith (1996)). Furthermore, there should be no difference in various self-administered modes (f.i., paper questionnaire or web-based surveys, Yan and Cantor (2019): 49) with respect to direct questions. ICT, however, shows a rather unconventional structure: the need of an instructional passage before the actual question. This difference in modes could be a key explanatory factor why results are different across studies.

One advantage of the ICT is the easy implementation compared to other indirect question techniques. There are even some indicators that suggest it is easier to understand than the Randomized Response Technique (Hubbard et al. (1989). Nevertheless, there are some limitations to this technique. For instance, as one part of the respondents receive a longer list than the other part, a systematic bias can not be ruled out. The results can be potentially unreliable. Holbrook and Krosnick (2010) tested for this bias directly, Tsuchiya et al. (2007) indirectly in online surveys along with De Jonge and Nickerson (2014) for face-to-face interviews and the results are inconclusive. Furthermore, some limitations exist concerning strongly prevalent behavior (Tsuchiya et al. (2007)) and such with a high sensitivity (Thomas et al. (2017)). A huge disadvantage is the fact that every sensitive item needs multiple non-sensitive items, and coming up with those is time-consuming. Another drawback of the ICT is the necessity of more participants caused by their division into two groups (Ahart and Sackett (2004)). Additionally, it is practically impossible to analyze the data at the individual level, which can be seen as major disadvantage (Yan and Cantor (2019)).

2.2 Crosswise Model

The origin of the second indirect question technique is the Randomised Response Technique (RRT, by Warner (1965)). By using the element of a randomized device e.g. dice or coin), the technique does not reveal the direct response to the sensitive question, but at an aggregated level (ibid.). The CM was theoretically first explored in 2008 (Yu et al. (2008)), but the first empirical study was only until later (Jann et al. (2012)). The method is thus significantly younger than the ICT and is deemed to be a more promising further development of the RRT (ibid., Walzenbach and Hinz (2019)). While former variations of the RRT offer a self-protective response and have the requirement of a randomization device, the CM can cope without them. Other than having to count the items, respondents of the CM are exposed to two questions at the same time and have to indicate whether the answers to those questions are the same or different (Yu et al. (2008)). One of the question is the sensitive one, the researchers care about to find out. The other question is unrelated and a non-sensitive question. It is chosen such, that the probability of the answer is reflected by a known distribution. The non-sensitive question could be for example about the birthday or birth month of the participants.

Given the same scenario as before (the sensitive item is about plagiarism and the respondent group are students), the following CM can be asked (based on Jerke, Johann, Rauhut, and Thomas (2019, Forthcoming) as well as Jann et al. (2012) and adapted to the purpose of this thesis):

`For this question, you are asked two questions in one block. Please start thinking about how you would answer each question individually (either yes or no), but do not write this down. Depending on how you would answer these two questions, please indicate whether the answer is (A) or (B) following the instructions below:`

`If your answer to both question is no or the answer to both questions is yes, please indicate this by selecting answer (A). If your answer to one of the questions is yes and it is no to the other, please indicate this by selecting answer (B).`

- `Is your mother's birthday in January or February?`
- `When writing your master thesis, have you intentionally adopted a passage from someone else's work without citing the original?`

`What is the answer to both questions?`

A) `Both questions yes or both questions no`
B) `One question yes and the other one no`

Respondents have to a tick either A or B, so to express that the answer to both is the same (yes or no) or different (one yes one no). Therefore, admitting to plagiarism on an individual level is again anonymous.

A requirement of the CM is choosing a non-sensitive question wisely. The distribution for that variable must be known, unrelated to the sensitive item and uneven, meaning not equal to 0.5 (Jann et al. (2012): 36). Furthermore, the sample size can not be too small, as the technique will not work.

The prevalence of the sensitive item can be estimated (Warner (1965), Yu et al. (2008)) by

$$p = \frac{pz - 1 + \frac{n'}{n}}{2pz - 1}; \text{ with } pz \neq \frac{1}{2},$$

where pz is the known prevalence of non-sensitive item, n is the sample size, and n' is the observed number of response where both answers are the same.

2.2.1 Applications of the CM

The geographical spreading of the CM is not as far-reaching as with the ICT, simply because the CM is comparatively new. There are studies from Iran (Shamsipour et al. (2014)) to Switzerland (Höglinger, Jann, and Diekmann (2016)) and the topics reach from voting behavior (Waubert de Puiseau, Hoffmann, and Musch (2017)) to xenophobic attitudes (Hoffmann and Musch (2016)). For an overview see section 5.1.1.

There is not much variation in regard to the non-sensitive item. The topics are mostly addressing the birthday (of the father, mother, the respondent or a friend; e.g., Canan (2017); Hoffmann and Musch (2016)), the street or house number of a person (f.e., Höglinger and Jann (2018)), the last digit of the pin code of a credit card (e.g., Khosravi et al. (2015)), the last digit of the phone number of a specific person (f.e., Safiri et al. (2019)) or an individual's first name (Fateme or Zahra for female and Ali or Mohammad for male persons; Vakilian, Keramat, Mousavi, and Chaman (2019); Vakilian, Mousavi, Keramat, and Chaman (2016)).

The results of CMs compared to other techniques are mixed. In some studies it performs better (Hoffmann and Musch (2016)) and in some it does not ((Coutts et al., 2011), Walzenbach and Hinz (2019)). The validation process is the same as with the ICT, the technique with a higher prevalence works better. This assumption will be discussed further.

2.3 More-is-better Assumption

In the research on indirect question techniques the standard measure to validate a method is the more-is-better assumption (Höglinger and Jann (2018)). Researchers use the results of direct questions (DQ) as a baseline. The underlying assumption is that respondents tend to underreport socially undesirable and overreport socially desirable behavior when asked directly (false negatives). Hence, an ICT or a CM study is more valid when it results in more (or respectively less) confessions than a DQ. Hence, a higher (or respectively lower) prevalence rate of the sensitive item produced in the ICT or CM as compared to DQ implies that the indirect question technique is better. Compared to other techniques and modes (e.g., Randomized Response Techniques), the ICT produces more valid results ((Lensvelt-Mulders, Hox, Van der Heijden, & Maas,

2005a), Tsuchiya et al. (2007), Sheppard and Earleywine (2013)). In the end, studies have shown mixed results on whether the ICT works better than direct questions. And while the more-is-better assumption is often legitimate, research on the CM shows it is not infallible (Höglinger and Jann (2018)). Research found unlikely high values and concluded that the CM could wrongfully artificially raise them (Walzenbach and Hinz (2019)), Höglinger and Jann (2016) were the first ones to actually investigate this phenomenon in a working paper (published Höglinger and Jann (2018)). The researchers showed, that the more-is-better assumption fails to consider false positives, i.e., respondents who admit something that they have never done (Höglinger and Diekmann (2017)) and that the CM in fact brings out these false positives. Höglinger (2016) investigated this further and proved the suspicions were justified (diseases with a spread close to 0% had suddenly a prevalence of 5-8 % in the CM condition; ibid: 94ff). To summarize, as there are no comparable alternatives the more-is-better assumption (or respectively less-is-better) still is the state-of-the-art to validate indirect questioning techniques.

3 Difficulties in Asking Sensitive Questions

Researchers design questionnaires in order to obtain true and unbiased answers to questions. The previous chapter described how researchers use indirect question techniques when they expect a high social desirability bias. Another frequent complication in questionnaires are non-responses, often caused by the cognitive burden of respondents who address the questions. This factor is not negligible in indirect questioning techniques as their unconventional style increases the cognitive burden of respondents. Therefore, this chapter first summarizes existing meta-analysis on the ICT and CM technique and how they fail to take the cognitive burden into account. Afterwords I discuss crucial characteristics of questions and their role with respect to the cognitive effort of respondents (e.g., how the the length of the question influences the capability to comprehend it). Finally, the importance of phrasing the question is explained and discussed.

3.1 State of the Art and Research Gap

A meta-analysis has been done on the original RRT (Lensvelt-Mulders et al. (2005a)) in which one of the most interesting conclusions is their verdict that the more sensitive the topic in question, the better the performance of the RRT. But as the researcher fail to explain most of the residual variance across studies, they conclude the RRT is not really controllable by researchers. It would be interesting to investigate how the presumably better version of the RRT, the CM, works in that respect. Up to this date there are no published meta-analyses on the CM, though there is one creation with the focus on the collocation of the sample of CM (Thomas, Schnell, and Noack (2019)). And while researchers take the country into account when analyzing the CM, to the best of my knowledge this has not been done before concerning the ICT.

There are a few meta-analyses on the ICT, as the technique has existed for longer and thus was applied more often. An analysis from 2007 compares the ICT with other techniques but many things have changed and multiple new studies were conducted since then (Tsuchiya et al. (2007)). The most recently published analysis is designed as a review (Hinsley, Keane, St. John, Ibbett, and Nuno (2019)), which concludes that researchers need to understand this method better in order for it to produce valid

results. Their focus is clearly on researcher's comprehension and survey designs. While there are some other meta-analyses in progress (Junkermann, Wolter, and Ehler (2019); Blair et al. (2018)), all fail to take a basic element into consideration: the instruction and how the design of the instruction affects the success of the method. The current study aims to shed some light on this research gap.

3.2 Cognitive Burden of Answering Questions

Before explicit characteristics can be investigated, the cognitive burden has to be explained. In order to answer a question, respondents have to go through four cognitive stages following the traditional model of cognitive mechanisms (Tourangeau (1984); Tourangeau, Rips, and Rasinski (2000)): Comprehension, Retrieval, Judgement and Response.

Comprehension. There are two main complementary approaches to study comprehension. The first one understands comprehension as a top down process of recognizing the pattern of the question and then identifying every piece in the pattern. The second bottom up approach understands comprehension as using prior knowledge and forming a coherent mental image (Tourangeau (1984)). Hence, the first stage includes attending the question and the accompanying instructions in addition to allocating meaning to it. The last key step to comprehending the question is then identifying the intent behind it (Tourangeau et al. (2000): 23f).

Retrieval. The second stage contains recalling important information from long-term memory. Respondents have to go back in their autobiographical memory and search for the incident or trait in question. The phrasing of the question which initiates this search can have a huge effect on the accuracy and completeness of this process (Tourangeau et al. (2000): 9f).

Judgement. After comprehending the question and retrieving the important information, respondents then make a judgment based on the retrieved memories. Although sometimes they answer based on the general plausibility of a response rather than real events (ibid.).

Response. In the final stage, judgment is translated into the response categories of the question. Ideally memories are selected and reported in an answer (ibid.).

There is a difference between actual behaviors and general attitudes with respect to retrieving autobiographical memories and forming an opinion about them. For a behaviour it is necessary to go back in time and recall an actual event. Instead, for an attitude the best scenario is: respondents have a preformed opinion on the topic in question and they are just waiting to be offered the possibility to express it. On the other end of the spectrum, respondents have no opinion whatsoever to the item in question. In between respondents might have an idea or loosely connected thoughts which still have to be combined (Tourangeau (1984); Tourangeau et al. (2000)). To summarize, answering questions in a survey requires respondents to invest a great deal of cognitive effort for little or no apparent reward (see ibid., Krosnick (1991)).

Indirect question techniques are cognitively more demanding as direct questions, since respondents need to evaluate multiple factors, consider numerous questions and cumulatively answer in one individual response. Therefore, it is common to introduce the question to respondents with a proper instruction. Whether respondents understand the question is, therefore, not clear and has already been subject to various research (see Jerke et al. (2019, Forthcoming); Krosnick (1991); Ahart and Sackett (2004); Coutts and Jann (2011)). As already discussed before, ICT-researchers experimented with the length and content of instructions. Giving a full briefing of the technique led to a higher tendency of admitting the sensitive trait (see chapter before and Dalton et al. (1994): 821). More explanation can lead to a better comprehension of the technique and therefore a higher prevalence rate of the sensitive item. At the same time, going through the process of how the method provides privacy and accordingly emphasizing anonymity over and over again also highlights the high interest of researchers in the sensitive item, resulting in the scenario where more respondents admit something that they never did. The underlying effect is called *acquiescence*; respondents agree independently from the question asked because they try to please alleged expectations of researchers (Bogner and Landrock (2015), Ahart and Sackett (2004)). So researchers then decided to cut or shorten the introduction and that also led in some cases to a better performance of the ICT (Ahart and Sackett (2004), Tsuchiya et al. (2007)). This again can have

multiple reasons. One of them is that the respondents did not fully comprehend the method and, thus, guessed the answer (Hoffmann et al. (2017)). Studies show that all indirect question techniques are found to be less comprehensible compared to DQ (Hoffmann et al. (2017)). Especially the CM seems to be more demanding than the ICT (Krosnick (1991); Jerke et al. (2019, Forthcoming)). Khosravi et al. (2015) prove a lack of understanding by people participating in questionnaires and ask if this is a result of the complexity of the instruction or the extraordinary answer categories, which in turn can hinder comprehension (Lenzner and Menold (2015): 1).

There are plenty of steps between asking the sensitive question and receiving an actual answer and all of them can be confronted with possible errors. The four stages by Tourangeau et al. (2000) are optional cognitive tools, dependent on other circumstances, for instance on how accurate respondents want their answer to be or how quickly they need to answer (Tourangeau (1984), Tourangeau et al. (2000)). There are different theories addressing at which point of the four stages such a bias occurs. Holtgraves (2004) hints at an evasion of the retrieval and integration steps, as some respondents answer what seems most desirable without taking personal occurrences into consideration. Another possible source is a more positive self-image combined with a confirmatory memory search (ibid.; Tourangeau and Yan (2007)). Further, the distortion can occur right before the response in order to avoid embarrassment (Tourangeau et al. (2000)). The most basic way to improve reports is to ensure that respondents have sufficient time to search for the memory at issue. Furthermore, it is important to motivate the respondent to invest the necessary effort and reduce the problems that might be encountered on the way (Cannell, Miller, and Oksenberg (1981)).

While the current study cannot investigate how much time was given or if the motivation for respondents was enough, it examines the strategies researchers chose to reduce problems of respondents, which manifest in characteristics of the appearance of the question.

3.3 Characteristics of Asking Questions

There is a vast amount of literature that provides guidance to researchers on how to ask questions in order to enhance responding (e.g., Bogner and Landrock (2015); Porst (2009); Lenzner and Menold (2015); Bradburn et al. (2004); Lenzner, Kaczmirek, and Lenzner (2010)). Researchers emphasize the need to avoid unfamiliar terms in the phrasing of the question (Lenzner and Menold (2015); Lenzner et al. (2010)). Changing a word to a more familiar one, as in *«booze»* instead of *«liquor»*, led to a 15% rise of reports (Bradburn et al. (2004): 104ff). Another psycholinguistic feature is the number of times a word occurs within a text. High frequently words require less processing time while reading and are, therefore, easier to understand than low frequently ones (ibid.).

When it comes to the length of a question, researchers are not in complete agreement on whether short or long is better. Some researchers advise to avoid long and complex questions (e.g. Porst (2009): 95; Lenzner et al. (2010)) in order to minimize respondents' effort. A common scenario would be a short question which leads to a quick ending and thus consistently satisfies respondents, as there is no need to spend too much time upon it. Such participants will probably also consent to future questionnaires. If a question, howerever, is too short, it might be imprecise or might include complex sentences which are complicated syntactically and therefore can hinder a smooth and easy reading and response (Tourangeau et al. (2000)). This can be prevented with a longer question. Another benefit of a longer wording is, that it gives more context which can be helpful to understand the intent behind the question (Tourangeau (1984)). Also, the longer the question the more it allows the researcher to ease into the topic (Seibert (2019)). Some research shows that long questions (with familiar words) increase the likelihood of reporting socially undesirable behavior (Bradburn et al. (2004)), while others deny the effect (Seibert (2019)).

In the case of indirect methods, longer instructions are not rare along with further explanations and examples on how to deal with the uncommon structure. Lots of time and space get lost while explaining the unfamiliar technique, the cost for respondents rise and the chances to drop out of the survey or never fill out a questionnaire again increases. The objection of the current meta-analysis is to shed some light on this trade-off every researcher has to ponder: On one hand, a longer question gives more

context and possibly even adds an extensive explanation on the mathematical process behind the question. Long questions also give way for examples, in order to demonstrate how to answer the question. On the other hand, a short introduction and thus a quick ending of the survey requires less time on the respondents behalf. Therefore, it is essential to look at indirect question techniques as a whole package, and investigate the cognitive burden of the respondents. Specific characteristics of the instruction can have an impact on the burden and thus on the performance of the method. So far characteristics of the instruction have not been addressed by prior research in that form and this poses a research gap.

4 Meta Analysis Design

This chapter captures the strategy for the meta analysis, the explanation of how the sample of this study has been gained, followed by an explanation of the coded variables. A meta-analysis is a synthesis of results of different studies (Lensvelt-Mulders et al. (2005a): 330) and allows insights to specific topics and how they have been researched. This thesis followed the six steps of an meta analysis based on Field (2010: 666): Beginning with an intensive literature research, defining inclusion and exclusion criteria, considering the effect size, calculating basic and further analysis till writing it down.

4.1 Compiling the Bibliography

In the first step, a bibliography was complied. For that reason I searched in the Web of Science and Google Scholar using the words and abbreviations 'ICT', 'Item Count Technique', 'Unmatched Count Technique', 'UCT', 'List experiment', 'list experiment + sensitive', 'list randomization', 'block total response method', 'unmatched block count', 'Crosswise *', 'Crosswise model' and 'sensitive + question'.

Furthermore, I consulted the key paper of the Item Count Technique (Droitcour et al. (1991), and searched through all papers citing it. The same procedure was repeated with the Crosswise Model, and two key paper (Jann et al. (2012) along with Yu et al. (2008)) were found further all paper citing them were checked. Finally, I also studied the working paper by Blair, et al. (2018), and went through their list of ICT studies. This lead to a first sample size of 384 papers.

4.2 Criteria of inclusion and exclusion

The first criteria were to only include published papers from the year of the origin of the corresponding technique, up to July 2019 which report a study. The threat of a publication bias is undoubtedly high, especially when testing a new method but there are plenty of studies that show that the techniques do not work, also non significant results are published. Furthermore, Junkermann et al. (2019) showed with their data which is similar to the data in this thesis that there is no real threat of publication bias.

After a screening of the initial sample of 384 papers, 129 theoretical papers without a study where excluded. I read 255 papers and had to exclude 98 because of the following reasons: (i) not an actual ICT or CM, (ii) no empirical evaluation study (e.g. simulation studies), (iii) unpublished manuscripts, (iv) foreign language (study and publication), (v) different versions of the methods, (vi) refers to the same study as another paper, (vii) only non sensitive questions were asked. This lead to 157 eligible papers, including 160 ICT and 33 CM studies. A comparison of the prevalence rates of the direct question and ICT or CM was another inclusion criteria for the analysis. Thus, 60 studies had to be excluded so the final sample contains 105 ICT and 28 CM studies. For a better overview see Figure 1.

The first idea was to take every study with a questionnaire into account and include all instructions in the full analysis. In order to do this I contacted in March 2019 79 researchers and asked for permission or a copy to the original questionnaire, so the instruction can be extracted from there. After more than one reminder and multiple weeks passing only 26 responded positively. Compiling all the information and complying with the deadline turned out to be impossible for this thesis, thus a criteria for the full analysis was whether the study was conducted originally in English and whether the original instructions were reported. Another reason for this restriction was to make the instructions more comparable.

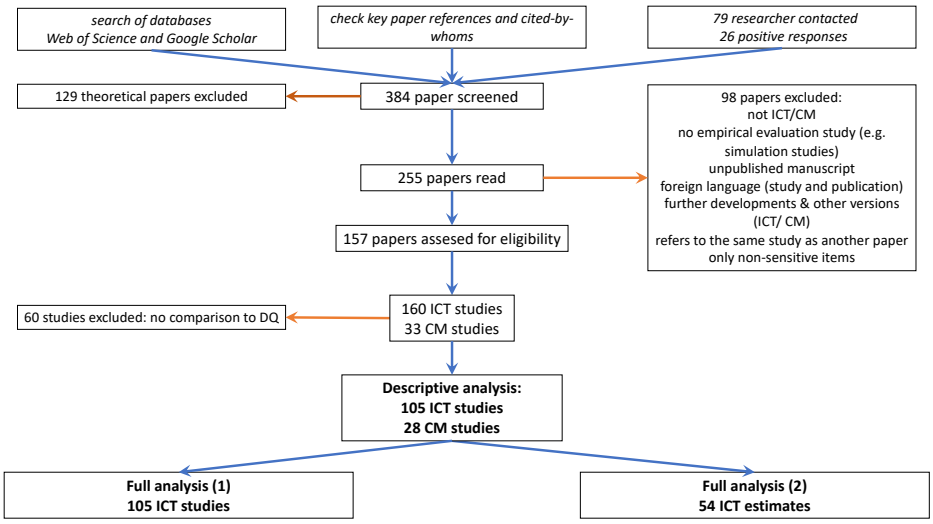


Figure 1: Procedure while compiling the bibliography. Results of descriptive analysis are presented in 5.1 and 5.2. For the full analysis (1) the variables are coded on method-level, for full analysis (2) the variables are coded on item-level.

4.3 Coding

For each method in each study, the following variables have been coded on the method-level (each question technique in an article equals one observation): (i) performance of the technique, (ii) country and language, (iii) sensitive topic, (iv) direction of the social desirability bias, (v) survey mode, (vi) pretest, (vii) number of non-sensitive items, (viii) context of items and (ix-xiii) five instruction variables that will be discussed in the following paragraphs. In the last step of the analysis all variables of ICT studies were additionally coded on the level of sensitive items (every sensitive item in a question in a paper equals now one observation).

Dependent variable. The effect size determines the performance of the indirect question technique. I choose not to use the raw mean difference (like other studies Junkermann et al. (2019)) between the indirect method and a direct question, as not every paper reports this with all detail, hence the sample size would be smaller. It was a trade-off, loosing information about the study or loosing studies and I decided to go with the first one. Following the more-is-better assumption (e.g. Umesh and Peterson (1991); Lensvelt-Mulders et al. (2005a); Thomas et al. (2017); Krumpal (2012)) higher descriptive estimates of the CM or ICT indicate better results. »Better«in this context refers to the case when the indirect question technique is superior in capturing the socially undesirable behaving or holding the respective attitude. If the item is socially desirable, the less-is-better assumption was used respectively. Even though there have been studies that show that following these assumptions blindly can be misleading (Höglinger and Jann (2018); see Section 2.3), there are no alternatives in the state-of-the-art, so this thesis also follows them, as have other meta-analysis before (Lensvelt-Mulders (2008), Hinsley et al. (2019)). For the performance, the effect size is coded as 1, if the special question technique over-performed, id est. yield a higher prevalence rate compared to the direct question. Everything else has been coded with 0. Hence, if the direct question and the indirect question technique generate the same prevalence, or if the DQ performed better than the other techniques the effect size is 0. The reason for this aggregation of information is, that the ICT and CM are often augmented as supposedly better than direct questions, and better is refering to a higher prevalence rate. To sum it up, the performance (X) is 1 if the prevalence (P) of a

socially desirable behavior is higher in the DQ-treatment than with the ICT or CM, but also lower if the trait is a socially undesirable one.

Socially desirable:

$$X = \begin{cases} 1, & \mathrm{P}_{DQ} - P_{ICT/CM} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Socially undesirable:

$$X = \begin{cases} 1, & \mathrm{P}_{ICT/CM} - P_{DQ} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Other Variables

Countries and languages. As the places, where the studies have been conducted vary hugely, country and language are coded but are only used for the descriptive overview. The ICT has been applied in the following regions: Afghanistan, Africa, Argentina, Austria/Germany/Switzerland, Brazil, Bulgaria, Cambodia, Canada, Chile, China, Colombia, Costa Rica, Egypt, Ethiopia, Ghana, Guatemala, Honduras, Hungary, India, Iran, Japan, Kenya, Korea, Lebanon, Liberia, Madagascar, Mexico, Netherlands, New Zealand, Nicaragua, Norway, Philippines, Romania, Russia, Senegal, Serbia, South Africa, Sri Lanka, Tanzania, Turkey, Uganda, Ukraine, United Kingdom, United States, Uruguay, Vietnam, Zambia, Zimbabwe.

For the regression analysis, countries are separated based on their status of democracy. I chose to split the countries based on their democracy status, as splitting them into whether they are western or eastern was too unspecific. The democracy level can have an influence on various stages of a study, for instance phrasing the sensitive item, designing the survey, getting approval to conduct the study, etc. On one side conducting a survey in authority countries could be linked to greater bureaucratic hurdles and therefore produce more thought-through designs which work better. On the other side a greater wish of respondents living in authoritarian countries to meet expectations of researchers is imaginable. To explore this further, I used the democracy index of the Economist Intelligence Unit, which collects data on electoral processes, pluralism, civic rights, government, political participation and culture. Their index provides variable manifestations about the level of democracy, which are summarized to either a democratic country (1) or authoritarian government and hybrid regime (0)[2].

Topics of the study. As the sensitive topic is central for the techniques, the issues researched within the studies are also coded. In order to broadly give a descriptive overview about all fields, I used the following categories: Abortion, academic dishonesty, affirmative action, attitude toward a religious group, attitude toward farmers market organization, attitude toward immigrants, attitude toward presidential election, attitude toward women as leaders, auctioneers behavior, blood donation, breastfeeding, bribery, corruption, delinquency and crimes, female genital cutting, flowers (orchids),

[2] https://infographics.economist.com/2019/DemocracyIndex/ last accessed on 30.09.2019

health and diseases, hunting, LGBTIQ, loan use, media, military, MTurk motivation, political receptivity, political views and trust, public service motivation, sexual behavior, substance abuse, tax evasion, unemployment benefits, unintended pregnancy, vote buying and voting behavior. These topics are summarized on a different level into two categories: attitude (coded as 1) and behavior (coded as 0) for the regression model.

Socially desirable vs. undesirable. The attitudes and behavior in question do not say anything about the direction of the social desirability bias. This code represents if the sensitive question is regarded as socially undesirable (coded as 1) or socially desirable (coded as 0).

Survey mode. The data collection methods are coded as a condition within a study to the following: Face-to-face interviews (chosen as reference category, as they provide the least of privacy), self-administered questionnaires (SAQ, e.g., pen and paper questionnaires), forms of telephone interviews (e.g. computer-assisted telephone interviewing) and web-based (online questionnaires). This distinction has also been made by other scholars (e.g., Lensvelt-Mulders (2008); Wolter and Laier (2014a)).

Pretest. If the researchers conducted and reported a pretest the variable is coded 1. If no information about a pretest is in the article, supplementary material, appendix, or any other source online, then the code is 0.

Item Count Technique. The following two variables are coded specifically for ICT studies. **Number of non-sensitive items:** The number of non-sensitive items in the long list of the item count technique are coded: 3, 4 and above 5. **Context of items:** This code refers to the case, if the non-sensitive and sensitive items are in the same range of topics, then the variable is coded 1.

Instruction variables. The following five codes refer to potentially crucial characteristics of the instruction, which in turn can have an impact on the performance. An instruction is defined as everything preceding the actual question. With the ICT it is the text proceeding the items and with the CM it is the text occurring before the two questions are asked. **Statistical explanation:** This code refers to the implementation of a statistical explanation how the technique works. Using one is coded as 1, refraining

from an explanation is coded as 0. **Use of an example:** I also coded if the instruction uses an example (1) to explain the technique or not (0). **Sentence structure:** The sentence structure is as discussed before a hint on the complexity of a text and, therefore coded as a ratio of words per sentences from 0–1. **Frequency of words:** By calculating the uniqueness of words and dividing it through the number of words, the variable repetition of words is coded from 0–1, where 0 means the instruction is full of word repetitions and 1 the instruction is full with unique words. **Number of words:** As an indication to the length of the instruction, the number of words were counted. While the first two content-related variables are coded manually, I automated counting words, the sentence structure and uniqueness of words (technical variables) using the natural-language-processing toolkit of Python.

5 Results

The resulting data set is compiled of 160 ICT studies and 33 CM studies. The sample size will differ, as the results of the indirect question techniques are not compared to DQ in every case and some studies refrained from reporting every information. Firstly, the results of the CM are presented. The main variables of interest are partitioned by the performance and descriptive results are shown. The next section shows the spread of the ICT and its performances so far, followed by a descriptive analysis. This is followed by the results of the regression models. Analyzing 28 CM studies ($N_{underperformed} = 12$, $N_{overperformed} = 16$) would be inconclusive, as the necessary sample size for the regression model is higher (Backhaus, Erichson, Plinke, and Weiber (2016): 347). Therefore, only ICT studies are addressed in the further analysis results. The first model shows the basic analysis without any instruction measurements, they are added successively, first the content-related instructions variables and then the technical. Due to the hierarchical structure of the data first models are calculated with the sample on the method-level. The next model shows the full analysis on a different observation level, the variables are coded on the item-level.

5.1 Crosswise Model

The CM was conducted in five different languages, most often in German (15 times, in Austria, Germany and Switzerland), then Persian (11, in Iran), English (5, in United Kingdom and United States) and Spanish (1, in Costa Rica) along with Serbian (1, Serbia). The top three most common topics are substance use (5 times), health and diseases along with delinquency and crimes and academic misconduct (each 4) and voting behavior as well as sexual behavior (each 3). Other topics are: Tax evasion(2), attitudes towards immigrants (2), unintended pregnancy (1), bribery (1), blood donation (1) and attitude toward women as leaders (1). For a better overview of the countries see Figure 12, of the languages see Figure 13 and of the topics see Figure 14 in the Appendix.

Figure 2 shows the distribution of the performance of the indirect question techniques and the question of how many of the CM did actually perform better than the DQ (following the more-is-better assumption) is answered. In total 16 CM performed

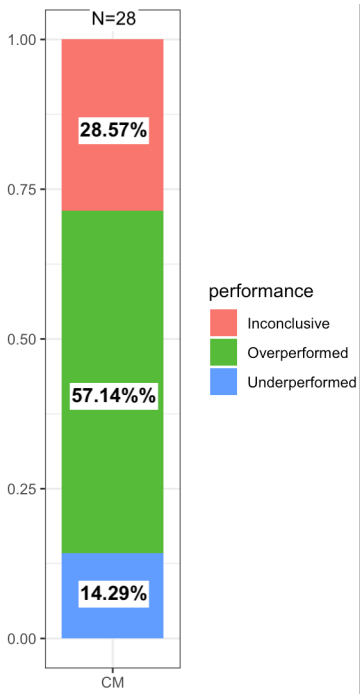


Figure 2: Roughly 3 out of 5 CM studies perform better than DQ

better than DQ (57.14%), four studies performed under or showed no difference to DQ (14.29%), eight (28.57%) showed inconclusive results and five did not compare the prevalence with those of DQ.

5.1.1 Descriptive Results

Democracy. From the 33 CM studies, 21 have been conducted in democratic countries and eleven in non-democratic countries, following the Economic Democracy Index, described in section 4.3. Figure 3 shows that from the studies in democratic countries that compared the CM with DQ (N = 21), 28.57% showed inconclusive results, in 4.76% the CM underperformed or showed no difference and in 66.67% the CM performed better than DQ. In non-democratic countries, the CM produced roughly 29% inconclusive results, roughly 29% higher and 43% lower prevalence rates of the sensitive item compared to the DQ.

Attitude vs. Behavior. In order to summarize the topics better, two categories are chosen: attitude and behavior. The CM was used for showing attitudes in a total of three cases, one of them was not compared to DQ and the other two performed better. Of the 29 behavior studies, four had to be excluded due to no comparison,

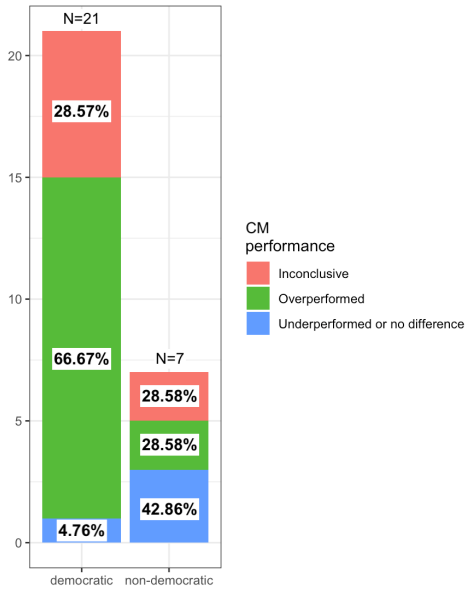


Figure 3: The CM seems to work better in democratic countries.

16% performed worse or showed no difference, 28% CM were inconclusive and 56% performed better than DQ (for further information, see Figure 15, in Appendix). The distinction of behavior versus attitudes in CM studies is not possible, as there are not enough studies with sensitive attitudes. Asking a sensitive behavior through a CM has been semi-successful, one out of two to performed better compared to asking the behavior in question directly.

Socially (Un-)Desirable For social desirability, there have been three studies with desirable and 19 studies with undesirable attitudes or behaviors among the CM. The CM seems not to work very well in socially desirable matters, as 66.67% underperformed and the remaining 33.33% showed only inconclusive results. For desirable matters, 10.53% of the cases the CM was beat by the DQ, 26.32% showed mixed results and in 63.16% the CM showed a higher prevalence than DQ did. A distinction between socially desirable and undesirable habits in in CM studies is not yet appropriate ($N_{desirable}$= 3). Socially undesirable items appear to work decently in the CM condition, in three out of five cases the indirect method performed better than the direct (see Figure 16 in Appendix for further information).

Survey Mode The studies which used CM have used different ways to ask questions (see Figure 4). While one study which used ACASI (audio computer-assisted self-

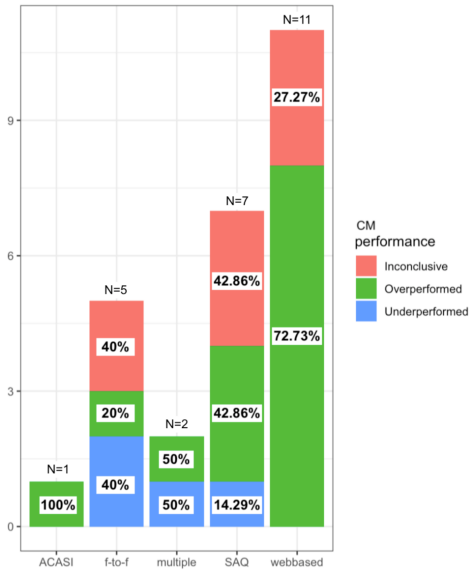


Figure 4: With a web-based survey mode the CM performs by far the best - 3 out of 4 cases yield higher results for the CM compared to DQ.

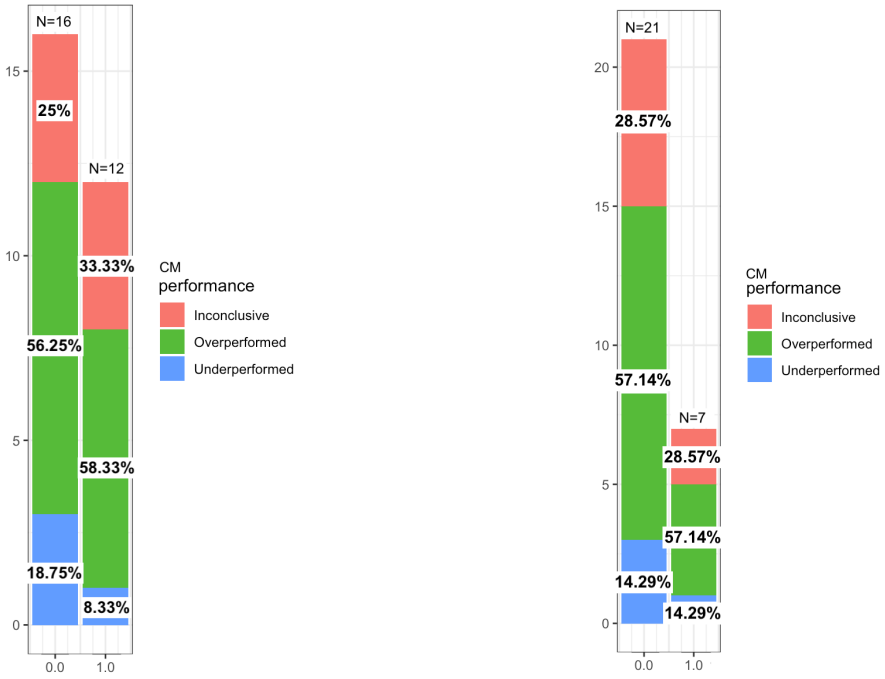
interview) overperformed, the other seven self-administered questionnaires (SAQ) were mixed: one underperformed (14.29%), three showed inconclusive results and three performed better than DQ (each 42.86%). Of the two studies, that use more than one mode, one over- and one underperformed. The face-to-face interviewed CMs show a tendency to perform worse, only one of the five yield a higher prevalence, while two had mixed results and two showed a lower prevalence than DQ. There are eleven web-based CM studies, of which three were inconclusive (27.27%) and promising eight over-performed (72.73%). To sum it up, the best performance of the CM was achieved when using web-based tools for the survey.

Pretest Of all considered CM studies, nine conducted a pretest and 19 did not perform or did not report it. Out of the pretested studies, five showed mixed results (55.56%) and four performed better than the DQ (44.44%). 15.78% of the 19 studies without a pretest indicate inconclusive results, 21.05% yield a lower and 63.16% a higher prevalence with the CM technique (see Figure 17 in Appendix for an overview). Contrary to all expectations, the CM performed more often better than the DQ, when not being subject to a pretest. This could be a trivial pre-finding with respect to the different sample sizes and should be analyzed further. It is also worth to mention that none of the nine CM studies, that have been pretested yield worse results than DQ

(five studies yield mixed results and four performed better).

Instruction: Content-related variables Twelve CM studies used statistical language or methods to explain the technique to their participants, while 16 declined to do so (see Figure 5a). 25% of the studies without a statistical explanation (N = 16) showed inconclusiveness, in 18.75% the CM underperformed and in 56.25% it overperformed compared to DQ. Of the users, 33.33% came up with mixed results, 8.33% showed the CM failing and 58.33% succeeding compared to the DQ. To summarize, it appears that the decision to include a statistical explanation has hardly any outcome on the performance of the technique. The only visible difference is the number of under-performed CM studies: it shrinks to half. But the sample size is not to be underestimated, further analysis are necessary for a more refined statement. Out of the 28 CM studies that compare the technique with DQ, seven use an example to explain it to the respondents and 21 do not use one (see Figure 5b). The performance is the same with both conditions: 28.57% came up with inconclusive results ($N_{example} = 2$, $N_{no\ exa.} = 6$), 14.29% show an under-performance of the CM ($N_{exa.} = 1$, $N_{no\ exa.} = 3$) and in 57.14% the CM yield a higher prevalence than the DQ ($N_{exa.} = 4$, $N_{no\ exa.} = 12$). In conclusion, the results of the usage of an example are similar to the use of an statistical explanation. There is no visible difference between using an example or not in CM studies except for the sample size. In order to be able make an assertion, further CM studies have to be conducted and analyzed.

Instruction: Technical Variables The following variables are chosen to display the instruction and take the complexity and length into account: (i) number of words, (ii) a ratio build to measure word frequency and (iii) the ratio of words per sentences. These variables are computed using a Python code, which can be found in Supplementary Materials. There are only four CM studies, that were originally conducted in English and report the instructions. While three instructions are rather long and extensive (144, 159 and 183 number of words), one study decided not to introduce the special technique at all: *«Please read the following two questions:»* with an additional *«Now select your response»* after the two questions (Roberts and John (2014), Supplementary p.4). The aforementioned article was excluded as this way of handling the introduction appeared to be an outlier. The sentences in the three articles are rather long (23, 24 and



(a) Explanation (b) Example

Figure 5: The performance of the CM compared to DQ appears to be not dependent from the use of an example or on having the method statistically explained beforehand or not.

26 words per sentence) and there are not many word repetitions (the ratio is close to 1 in all three cases). These are all indications to rather long and complex instructions. But as the sample size is too small, no conclusion can be drawn at this point.

5.1.2 Discussion

Of the initial 33 Crosswise Model studies, five had to be excluded. Roughly 60% of the remaining sample performed better (57.14%), while 28.57% showed inconclusive results and 14% performed worse or showed no difference. In the descriptive analysis the CM seems to work better in democratic countries (roughly 70% over-performing compared to 30 %) through online surveys. This is in line with other research Thomas et al. (2019). Whether asking a socially undesirable or desirable item or a behavior or an attitude is still to be explored. As the sample sizes alter significantly, no verdict can be drawn yet. Further, the performance of CM appears to be independent from the use of an example and a statistical explanation. The CM studies (N = 3) mark a mean of 162 (144-183, excluding the outlier) words per instruction, while having rather low frequent words (close to 1) and comparatively long sentences with a mean

of 24 words per sentence (23-26, excluding the outlier). These are all indications to rather long and complex instructions, which should be looked into more in-depth, but with a sample size of 3 there are not many conclusions that can be drawn at this stage. Furthermore, it appears that conducting a pilot before the main study has no influence on the performance, but it is hard to draw conclusions as it is possible and likely that more studies conducted pretests, but did not report it in the manuscript, supplementary or appendix (Hinsley et al. (2019): 312).

5.2 Item Count Technique

The studies using the Item Count Technique were conducted in 49 countries: most frequently in the US (69 times) followed by Austria/Germany/Switzerland (14) and Russia side to side with China (each 4). Then Nicaragua, Ethiopia and Argentina (3) and United Kingdom, Sri Lanka, Romania, Mexico, Lebanon, Kenya along with Bolivia (each 2). In the remaining 35 countries, ICT was only used once per country, the whole list can be found in the Appendix (see Figure 18). The most common languages for the ICT studies, are English (77 times), Spanish (16) and German (14), followed by African languages (7), Russian and Chinese (each 4), Swahili together with Arabic (each 3) and Romanian as well as Amharic (each 2). Furthermore, there are a lot of languages (in total 21) which appear only once, and can be found in Figure 19 in the Appendix. Furthermore, the topics vary with an emphasis on behavior and attitudes around voting. The most frequent topics of the ICT were the following in decreasing order (see Figure 20 in the Appendix): vote behavior (25 times), vote buying (21), delinquency and crimes (16), attitude toward immigrants (16), health and diseases (12), political views and trust (10), attitude toward presidential election (10), Military (7), LGBTIQ (4), Substance use, sexual behavior, public service motivation, attitude toward religious groups, academic dishonesty and abortion (each 3), female genital cutting, attitude towards women as leaders and affirmative action (each 2).

Of the 160 Item Count Technique studies, 55 did not compare the results with those of direct questions and therefore have to be removed from the sample. Of the remaining 105 studies, 31 (29.52%) showed mixed and inconclusive results, 32 (30.48%) performed under or showed no difference and 42 (40%) performed better than direct questioning.

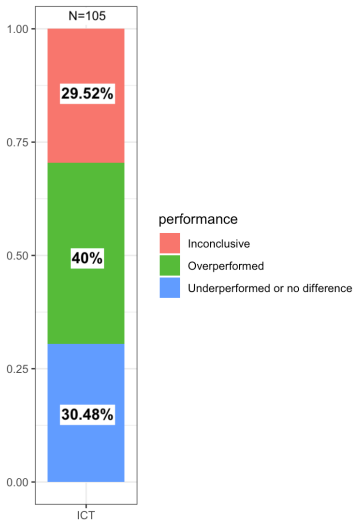


Figure 6: Two out of five ICT studies perform better than DQ.

5.2.1 Descriptive Results

Democracy From all ICT studies which have not been conducted in multiple places at the same time, 81 took place in democratic countries and 19 in non-democratic countries. In democratic countries one third performed better (33.33%), one performed worse (34.57%) and one showed inconclusive results (32.10%). In non-democratic countries (N = 19) the ICT performed in 73.68% better than DQ, 15.79% showed inconclusive results and 10.53% the DQ were equally good or better than the ICT. At the first glance, conducting the technique in non-democratic countries yielded better results but the sample size differs non negligible. This has to be analyzed further (see Section 5.2.2).

Attitude vs. Behavior Of the 39 ICT studies that had attitudes as sensitive items, 28.21% came up with mixed results, and equally 35.90% performed worse and better than DQ (see Figure 21 in Appendix). Asking about a behavior (N = 64) showed better results for the ICT, 42.19% performed better, 28.13% yield a lower prevalence or no significant difference to DQ and 29.69% showed inconclusive results.

Socially (Un-)Desirable From all ICT studies, 40 have to be excluded because they show no comparison to DQ and further 38 studies do not indicate whether the sensitive item is socially desirable or undesirable. From the remaining studies 15 cover a desirable attitude or behavior. 20% showed inconclusive results, 53.33% achieved a

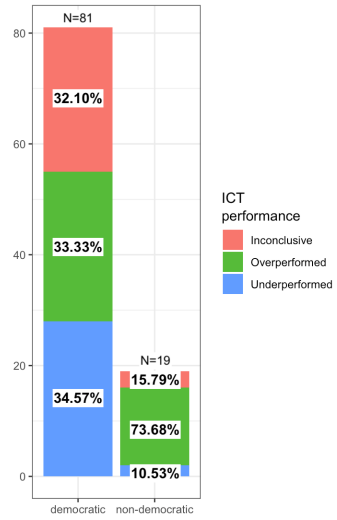


Figure 7: Democratic vs. Non-democratic Countries: ICT in on first glance in non-democratic countries.

lower prevalence than DQ and only 26.67% performed better. 65 ICT studies covered socially undesirable behavior or attitudes. Out of those, 26.15% reach inconclusive results, 29.23% performed worse and 44.62% performed better than direct questions. There are also not many ICT studies regarding socially desirable behavior. Out of them one out of every third study worked better. Undesirable items seems to work better, close to 45%, that means in nine out of 20 cases ICT performs better than DQ.

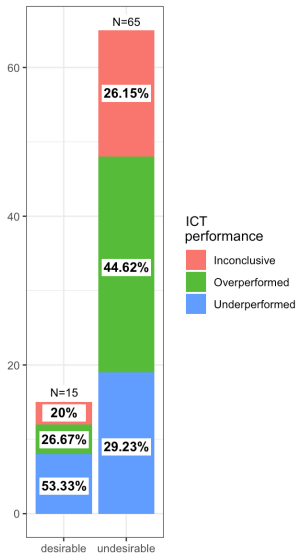


Figure 8: ICT studies with socially undesirable items perform better when compared to desirable.

Survey Mode Similarly to the CM, the ICT was also conducted once using ACASI (which performed worse than the equal DQ) and once through multiple modes (which came up with mixed results). For face-to-face interviews the following results were accomplished: 20% were inconclusive, 13.33% underperformed and 66.67% performed better than DQ. 13 ICT studies were self-administered, 58.85% of those showed inconclusive results, 30.77% performed worse and 15.38% better than DQ. 8.33% of the 12 telephone studies were inconclusive, comparing ICT to DQ 41.67% yield a lower and 50% a higher prevalence. Of the 44 web-based ICT studies, 16 showed inconclusive results (36.36%) only ten were better (22.73%) and 18 (40.91%) performed worse than the DQ or showed no difference. With the limited data that was available, face-to-face interviews appear to be the best solution, however, further analysis would be needed in order to draw conclusions.

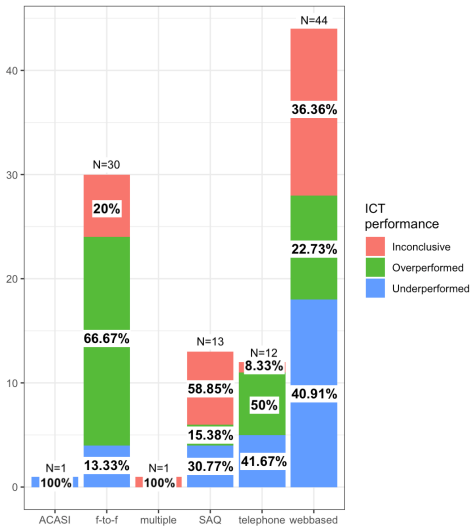
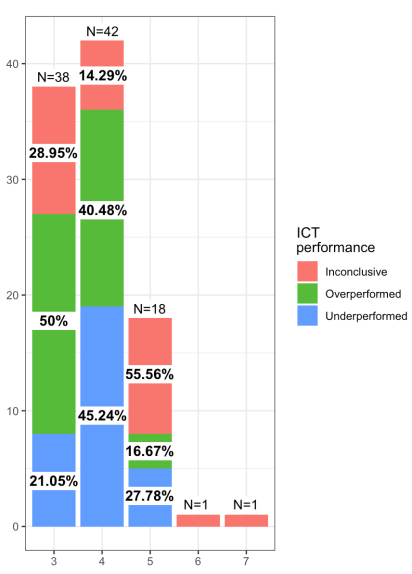


Figure 9: The ICT yields most frequently best results when asked face-to-face.

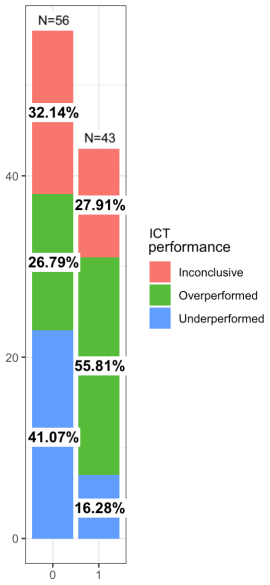
Pretest Of all ICT studies, 30 chose to conduct a pretest before the main study. In those, 36.67% report inconclusive results, 23.33% performed worse, and 40% had a higher prevalence than DQ. 75 studies were not pretested before or not did not report it. In these studies 40% performed better, 33.33% performed worse or same as than DQ and 26.67% showed mixed results. There is no significant distinction between pretested and non-pretested ICT studies, but this could also be due to very different sample sizes.

Non-sensitive Items and Context of Items There were 38 studies with three, 42 with four, 18 with five and each one with six and seven non-sensitive items (see Figure 10a). As the latter two showed only inconclusive results, the best choice seems to be three items: 50% of the ICT studies performed better and only 21.05% worse or the same as DQ (28.95% showed mixed results). Within the category of four non-sensitive items 40.48% yield a higher, 45.24% a lower prevalence and 14.29% inconclusive results. In the remaining category, only three ICT studies (16.67%) performed better than DQ, five worse (27.78%) and ten (55.56%) produced mixed results. The ICT seems to perform more often better than the DQ in most cases when having three or four non-sensitive items. Further analyses are required in order to distinguish between them. There are 56 ICT studies, in which the context of the non-sensitive and sensitive items is not equal or similar. Of those, 15 (26,79%) yield a higher prevalence than the respective DQ, 23 a lower prevalence (41.07%) and 18 (32,14%) showed mixed results (see Figure 10b). Of the 43 studies that chose a similar context for the items, 24 (55.81%) yield a higher prevalence in the ICT condition, seven a lower one (16.28%) and 12 came up with inconclusive results (27.91%). To sum it up, the ICT provides a better performance (more often a higher prevalence) when the non-sensitive and sensitive items have a similar contextual background.

Instruction: Content-related variables Only in five ICT studies researchers decided to use a statistical explanation (see Figure 23a in Appendix) of which three (60%) showed inconclusive results, one ICT performed better (20%) and one worse (20%) than DQ. Of the 100 studies that had no explanation, 28% came up with mixed results, 31% yield a lower and 41% a higher prevalence in the ICT treatment. There are not many ICT studies with an explanation, therefore, comparing the two conditions is not appropriate. Two out of every fifth ICT study without an explanation yielded a higher prevalence than the equivalent DQ. There are ten ICT studies with an example and 95 with no use of one (see Figure 23b in Appendix). Of the example-users, 40% came up with inconclusive results, 40% perform worse and 20% ICT perform better than direct questioning. 28.42% of the non-users generate mixed results, 29.47% have a lower prevalence and 42.11% have a higher prevalence in the ICT treatment compared to DQ. A direct comparison of the ICT studies is again not advisable as there are hardly any studies with an example.



(a) Number non-sensitive Items



(b) Context of items

Figure 10: ICT performs more often better when having 3 or 4 non-sensitive items. A shared context of sensitive and non-s. items leads to a higher chance of a better performing ICT from around a third to more than one in two cases - considering the sample sizes.

Instruction: Technical Variables Again, the variables were coded: (i) Number of words, (ii) a ratio build to measure word frequency and (iii) the ratio of words per sentences. For the ICT, the sample size is 38. The number of words in the instruction has a mean of 35 (range = 8 – 58). The ratio of frequency of words per instruction has a mean of 0.79 (range = 0.53 – 1). This ratio is reported on a scale from 0 (every word is at least twice in the instruction) to 1 (every word is unique, there are no repetitions in the instruction). The ratio of words per sentences has a mean of 14.79 and a range of 8 – 23. The instruction as well as the sentences seem rather short, and with high frequency word, thus, many word repetitions.

5.2.2 Results on Method-level

Table 1 shows the results of a binary logistic regression, the basic model in this thesis. Coefficients are reported together with the standard error in parentheses. The following variables are included: democracy of a country, direction of the social desirability, if the topic concerns an attitude or behavior, survey mode, the use of a pretest and the number of non-sensitive items along with the context of the items. Their influence on the performance of the ICT is calculated. The model accuracy indicators are in the

Variables	Model 1***		Model 2*	
	Coeff (sd)	*p*	*Coeff (sd)*	*p*
Constant	2.27 (1.37)	0.098	2.25 (1.37)	0.101
Democratic country	-1.48 (1.28)	0.247	-1.46 (1.27)	0.251
Socially undesirable	1.77** (0.89)	0.047	1.77** (0.89)	0.047
Attitude	-0.62 (0.84)	0.456	-0.64 (0.84)	0.446
Survey Mode (Ref.: F-t-f)				
Self-administered	-3.69*** (1.56)	0.018	-3.55*** (1.59)	0.026
Telephone	-0.70 (1.37)	0.608	-0.7 (1.37)	0.595
Online	-2.60*** (1.20)	0.030	-2.58*** (1.20)	0.032
Pretest	-1.13 (.90)	0.211	-1.08 (0.91)	0.236
Quantity non-s. I. (Ref.: 4)				
3 non-sens. Items	-0.52 (.98)	0.598	-0.53 (0.98)	0.589
>5 non-sens. Items	-1.34 (1.25)	0.282	-1.39 (1.29)	0.279
Context Items	-0.51 (.80)	0.524	-0.44 (0.82)	0.593
Explanation			-	-
Example			-1.22 (2.01)	0.558
N	*65*		*65*	
McFadden's Pseudo R2	*0.361*		*0.356*	
Count R2	*0.800*		*0.800*	
Log likelihood	*-28.189*		*-27.999*	

Table 1: Results of Logistic Regression on Method Level

acceptable areas (Backhaus et al. (2016): 317 for the R^2 and Backhaus et al. (2016): 315 for the likelihood).

In Model 1 a significant influence towards a good performance of the ICT over the direct question with the following variables can be witnessed: Asking a question with a social undesirable answer ($OR = 5.86$), choosing the survey mode of face-to-face interviews compared to self-administered ($OR = 0.02$) or web-based ($OR = 0.07$). Given the descriptive analysis the direction of the bias, which is increased when an undesirable trait is in question along with the decrease in SAQ and web-based surveys compared to face-to-face is expected like that. And while the literature is torn in the matter of the direction bias, the results concerning the survey mode are rather surprising and will be discussed later. Face-to-face interviews in comparison to telephone interviews do not result in significant differences ($p > 0.608$). Conducting the ICT in an undemocratic country preforms better than in a democratic, but the difference is not significant ($p > 0.247$). Asking about attitudes instead of behaviors decreases the chances of a better performed Item Count Technique compared to direct questions, but again the difference is not significant ($p > 0.456$). The same applies to the case of a conducted pretest before the original study: The chances of an overperformed ICT decrease with a pretest ($p > 0.211$). The difference between three, four or more than five used non-sensitive items indicates a better performance in the decision of using four but is not significant ($p_3 > 0.659$; $p_5 > 0.282$). Choosing the same context for sensitive and non-sensitive

items is highly significant when calculated alone ($p = 0.004$) but the significance is lost when all variables are added in the last step (Table 6: M 1.7 compared to Model 1 in Appendix).

In the next step the content related variables about the instruction are added (see Table 1). Model 2 (see Table 1) shows significant results with the same factors as Model 1: social undesirable items, which are surveyed through face-to-face interviews. The rest is unsurprisingly not significant. Social undesirable sensitive answers ($p > 0.047$) show a positive effect on the performance (compared to DQ). Conducting the survey through self-administered questionnaires ($p > 0.026$) or through an web-based mode ($p > 0.032$) decreases the probability of an over-performance compared to face-to-face interviews. About the other variables, which are not significant, only the direction can be interpreted cautiously. Conducting the research in a non democratic country might have a better performance of the ICT (*Democratic country:* $OR = 0.23$, $p > 0.251$), as well as asking about a sensitive behavior (*attitude:* $OR = 0.52$, $p > 0.446$) or renouncing on a pretest (*pretest:* $OR = 0.34$, $p > 0.236$). There seems to be a tendency toward the face-to-face interview compared to a telephone interview, but the difference is not significant (*telephone:* $OR = 0.48$, $p > 0.595$). There are indications that choosing four non-sensitive items compared to three (*3:* $OR = 0.59$, $p > 0.589$) or more than five (*more than 5:* $OR = 0.248$, $p > 0.289$) has a positive influence on the performance of the ICT, but there is no significant distinction. As for the sensitive item sharing the same context, the tendency points toward a dissimilar situational background (*same context:* $OR = 0.65$, $p > 0.593$). The addition of a statistical explanation had to be excluded due to the low number of observations (N = 3). It seems to be wise not to use an example when introducing the ICT (*example:* $OR = 0.296$, $p > 0.558$) but again, the difference is not significant. The second model did not improve in comparison to the first model (see Count R^2 values), thus the conclusion can be drawn that the two content-related variables do not explain the different outcomes across the studies.

The next step was to also include also the remaining variables about the instruction (see Model 3, Table 6, in Appendix). With this step, all significant values disappear. There are plenty reasons for this outcome. The inclusion of the variables about the instruction limits the sample to studies conducted in English-speaking countries which

Variables	Model 4	
	Coeff (sd)	*p*
Constant	57.27 (1584.242)	0.971
Democratic country	-	-
Socially undesirable	5.24* (3.09)	0.090
Attitude	3.38 (2.55)	0.185
Survey Mode (Ref.: F-t-f)		
Self-administered	-21.62 (1584.11)	0.989
Telephone	-6.44 (4.17)	0.122
Online	-22.351 (1584.11)	0.989
Pretest	-2.92 (2.91)	0.316
Quantity non-s. I. (Ref.: 4)		
3 non-sens. Items	-1.10 (1.71)	0.518
>5 non-sens. Items	-.079 (1.32)	0.952
Context Items	5.09* (3.01)	0.090
Explanation	-	-
Example	-	-
Words/sentences	-0.50 (0.35)	0.145
Word repetitions	-32.90** (16.39)	0.045
Number of Words	-0.24* (0.13)	0.076
N	*54*	
McFadden's Pseudo R^2	*0.115*	
Count R^2	*0.661*	
Log likelihood	*-36.020*	

Table 2: Results of Logistic Regression on Item Level

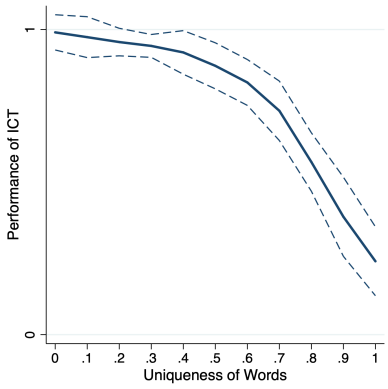
report the instruction; reducing in a size of N = 18. The results of this model are shown in the Appendix. Given this consequence, all variables were coded on the level of items instead of methods. This led to dissolving the group of inconclusive results, as every sensitive item is now coded as one observation. The new sample size is now at N = 54 (see Model 4, last column in Table 2).

5.2.3 Results on Item-level

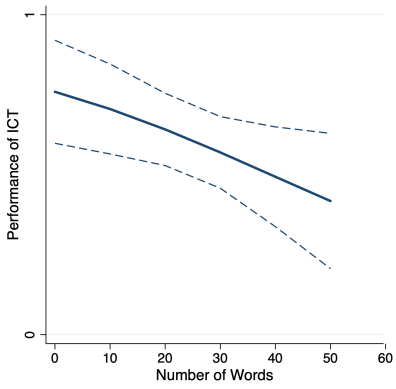
At first glance it is evident that one variable shows a significant coefficient at the 5 % level (uniqueness of words in the instruction) and three at the 10% level: socially undesirable, context of the sensitive and non-sensitive items and number of words. Choosing the same context when creating the non-sensitive items increases the probability of a overperforming ICT in comparison to its DQ (*context: $p > 0.090$*). The same occurs when phrasing a socially undesirable item instead of a socially desirable one (*$p > 0.090$*). Also, some parts of the instruction seem to have an influence on the performance: The shorter the instruction and the more use of words repetitions, the higher the probability of the ICT to perform better than DQ (number of words: $p > 0.076$; uniqueness: OR = 5.13e-15[3], $p > 0.045$). This is also illustrated in Figure **??** (in line with current standards I report marginal effects, computed according to the *«observed-value-approach»*, Hanmer

[3] 5.13e-15 = 0.00000156927

and Ozan Kalkan (2013)).



(a) Uniqueness of words: A decrease of the likelihood of the ICT performing better as the ratio increases. The more unique words, meaning the less repetitions of words in the instruction the less likely the ICT performs better.



(b) Number of Words: A decrease of the likelihood of the ICT performing better than DQ with as the number of words in the instruction increases.

Figure 11: Visualization of Technical Instruction Variables (ICT)

All the remaining variables have a possible effect on the performance, but show no significant difference. For a increased probability of a ICT yielding a higher prevalence, the item in question should be about an attitude ($p > 0.185$), not pretested before the main study ($p > 0.316$) and hidden within 4 non-sensitive items (>5: $p > 0.952$; 3: $p > 0.518$). The question should be asked in a face-to-face interview (SAQ: $p > 0.989$, $telephone$:$p > 0.122$, $web\text{-}based$: $p > 0.989$) and the instruction should not be constructed with very long sentences ($p > 0.145$).

Cautiously comparing the Models 1, 2, 4 shows similar directions, which should not be taken too serious, as the sample differs. But we can look at the directions, this can help further research. Choosing to conduct the indirect question technique in a non-democratic country, drafting the sensitive topic in question as a socially undesirable one (significant in all three models), carrying out the survey as a face-to-face interview, not running a pretest beforehand and choosing four non-sensitive items to hide the sensitive element in question increase the likelihood of a better performing ICT. While the first two models show that it seems to be wiser to ask a behavior instead of an attitude as an indirect question, this changes when the variables of the instruction are added and the sample size changes. Something similar can be observed when investigating the context

of the items and the number of non-sensitive items. Choosing four non-sensitive items without any context to the sensitive one seems promising for the technique without taking the instruction into account, and dividing the results of the inconclusive results, While there is no significant difference analyzable in regard of the higher numbers ($p > 0.563$), using the similar contextual background is as said before significant on the 10% level ($p > 0.060$).

5.2.4 Discussion

Of the 160 Item Count Technique studies, 55 did not compare the results with those of direct questions and had to be excluded. Of the remaining 105 studies, 40% performed better along with equally 30% worse or showed no difference and 30% inconclusive.

So far, there have been no meta-analysis of the ICT which take a deeper look at the country where the method was employed. Following the descriptive analysis, the method appears to work better in non-democratic countries (73.68% overperforming compared to 33.33%). Also, the results of the regression point into the direction that the ICT outperforms the DQ in an authoritarian government or hybrid regimes but the difference to democratic countries is not significant. Further research is necessary to explore this more. Nonetheless, different method of coding the countries is advisable, as dividing hybrid regimes and authoritarian countries, which are combined in the Economist Intelligence Unit Index, could give more insight to this.

A socially undesirable item performed better in the descriptive analysis than a socially desirable one (44.62% overperforming compared to 26.67%). This reflects also in the full analysis, the variable remains significant. This is also in line with literature, where previous research found these methods worked better when confronted with a social desirability bias in the same direction (see Chapter 3. and Bradburn et al. (2004)). While there is a difference in reporting behaviors or attitudes when it comes to retrieving memories and forming an opinion, no significant differences are to be found in this meta-analysis. Neither the descriptive part nor the regression show conclusive results; while the first two models point in the direction that asking about a behavior instead of an attitude increases the chances of the ICT performing better, this changes when the variables of the instruction are added.

All results also point in the direction that the ICT works most often best when asked in a face-to-face setting (descriptive: 66.67% overperforming compared to the rest, see Figure 9 and in the analysis see Table 1). While this finding contradicts the general literature, it is in line with another meta-analysis on the ICT (Hinsley et al. (2019): 316). General research on sensitive questions pursues the idea that this form of execution (face-to-face interviews) is contradicting to respondents feeling free to admitting. A reason for this rather odd result could be due to the comprehension and the way of how the respondents are confronted with the special technique. While respondents have to read the instructions in an online and self-administered survey, during a face-to-face interview, respondents hear the instructions. This could be interesting on more levels, as survey research always suggested a difference in hearing, reading (Tourangeau (1984)) or seeing for instance an illustration of items.

The second model did not improve in comparison to the first model, thus the conclusion can be drawn that the two content-related variables do not explain the different outcomes across the studies.

Conducting a pilot beforehand is undoubtedly advisably in any kind of indirect question technique, even though the results point in a different direction. It appears conducting a pilot before the main study has no influence on the performance, but it is hard to draw conclusions, as the sample size differs between the characteristics significantly. It is possible and likely that more studies pretested their version of the ICT, but did not report it in the manuscript, supplementary or appendix (Hinsley et al. (2019): 312).

In regard of the ICT specific variables, the number of non-sensitive items and contextual background, the following results were obtained: In the descriptive analysis the ICT performed best when the sensitive item hidden within three or four non-sensitive items (50% overperforming and 41% compared to the rest, see Figure 10a), which are chosen from the same context (55.81% overpreforming compared to 27%). Without taking the instruction variables into account, choosing four non-sensitive items without any context to the sensitive one seems promising for the technique. Even though the results of this thesis in regard of the length of the item list are ambiguous, all other characteristics being equal, longer lists offer more privacy but at the same time load a bigger cognitive burden on the respondent (Tsuchiya et al. (2007)), which is visible

though not significant in the current results. While no significant difference was observed in regard to the higher numbers ($p > 0.563$), using the similar contextual background becomes significant on the 10% level ($p > 0.060$). Thus, selecting the non-sensitive items from the same contextual background as the sensitive item results in a better performing ICT. This could be explained with the simple fact that the sensitive item does not stand out if the context is similar. A prominent sensitive item might raise suspicion.

Concerning the instruction variables, there are not many ICT studies which use an example or further explanations to explain the technique. The sample ($N = 38$) shows a mean of 35 words per instruction of 35 (8-58) with a frequency ratio mean of 0.79 (0.53-1), which indicates rather many repetitions of words per instruction. In regard to the structure of sentences, the mean of the words per sentence ratio is roughly 15 (8-23), which indicates rather short sentences. The possibly most interesting part of this study is probably also its contribution: a shorter introduction has a higher chance of becoming an overperforming ICT compared to the equivalent DQ *($OR = 0.79$, $p > 0.076$)* and so does using more unique words compared to word repetitions *($OR = 5.13e\text{-}152$, $p > 0.04$)*. This result reinforced the part of the scientific discourse which emphasizes to avoid long or complex instructions, as it only maximizes respondents effort. This can be interpreted as an indication on a bigger level: The less the experienced weight of a cognitive burden on respondents, the better the ICT performs. But only future research can find evidence on these hypotheses. Further, it is in line with research conducted on ICT: Ahart and Sackett (2004) found significantly higher ICT rates in the case where they gave their respondents instruction than in the condition without instructions, as it compliments the state-of-the-art with the aspect of short introductions.

6 Conclusion

Questionnaires covering sensitive topics suffer from an possible under- or overestimation of the true prevalence due to the desire to gain social approval. To overcome this social desirability bias, new and promising questioning techniques as the Item Count Technique and the Crosswise Model have been explored by researchers. Both techniques show promising results (Lensvelt-Mulders et al. (2005a); Tsuchiya et al. (2007) and see Chapter 2), but they come at a cost. Researchers have to make situational decisions and agree to a trade-off between more anonymity of participants through an unconventional question structure on the one hand, and shorter, less complicated and also less anonymous direct questions on the other hand. Difficult to understand questions can lead to different interpretations, incorrect answers, or satisficing (the wish of respondents to provide satisfying rather than optimal answers) (Lenzner et al. (2010), Krosnick (1991)). This meta-analysis sheds some light on previous implementations of the ICT and CM techniques and provides best-practices for situations in which they are more likely to work.

In this thesis I conducted an extensive web search to find all published studies and coded their differences and salient characteristics for the future analysis. Different applications of the ICT and CM have been compared and I coded possible reasons why the results differ across studies. To include the instruction in the analysis, I coded two substantial (the use of an example or a statistical explanation) and three technical (number of words, frequency of word repetitions and further sentence complexity) features for the analysis.

To conclude, both indirect question techniques indicate mixed results. The Crosswise Model yielded in three fifths of the studies a higher prevalence of the sensitive item compared to direct questions. The Item Count Technique performed in two fifths of all studies better than direct questioning. Although at the first glance a better performance of the CM can be witnessed, but the current sample consisted of three times as many ICT studies than CM studies. However, only applying those methods is no guarantee for success, it rather matters on how and in what setting exactly they are implemented. The CM tends to work better in democratic countries, while the ICT yields better results in non-democratic countries. Furthermore, the CM shows a tendency to work

better in online survey settings, while the ICT works best in face-to-face interviews with items chosen from the same contextual backgrounds and a socially undesirable sensitive item.

To the best of my knowledge, this thesis is the first work analyzing the characteristics of instructions on the success of the indirect questioning methods. On the example of the ICT it is shown that specific characteristics of the instruction have a significant influence on the methods success. Instructions with too many words seem to have a negative influence as well as the number of word repetitions (a common proxy to estimate how complicated a text is written) also significantly influences the outcome. These results have been shown on the ICT but they could be a valid indicator for the success of other indirect questioning techniques as well, as they reinforced the part of the scientific discourse which emphasizes to avoid long or complex instructions, as it only maximizes respondents effort. But only future research can find evidence on this.

6.1 Limitations and Future Work

Taking the decision to code the effect size as binary dependent variable led to a higher sample size but limited the analysis as the magnitude of the effect is removed. Future work can choose to calculate a three-level weighted regression model, with the first level being the studies, the second the method and on the third the items, to reach additional and more refined results. This suggestion comes with the premise that the number of respondents has to be coded and added to the model and the effect size is not binary but e.g. the real difference in prevalence. While this approach probably reduces the sample size (due to missing reports of those number), it also makes it possible to calculate individual validation studies and see if there are indications for differences between studies and conditions (in regard of sampling error, residual error terms at study level and condition-within-study level; for more insight see Lensvelt-Mulders et al. (2005a):332). In order to be able to draw more refined conclusions about the Crosswise Model, the data can be coded on the item level. This procedure increases the sample size and possibly allows for a regression analysis.

Another possibility for future work is to consider different codes of countries and topics on a more substantial level as well as different content related instruction variables.

Also, some new codes can be considered. As already discussed, the way how respondents are confronted with the technique (e.g. reading the instruction or hearing explanations or seeing the illustration of the items) might give more insight on the performance of the method.

Further an additional variables concerning the sensitive item is advisable. Following Lensvelt-Mulders et al (2005a: 329) procedure, experts and amateurs can be asked to rank the items on a scale from 0 (no inclination toward social desirable answering should be expected) to 4 (the researcher can hardly expect an honest answer to this question). Also inspired by the same researchers, adding a variable coding the quality of the data can bring some new insights. The code include sample size adjustments, whether researchers investigated a convenience sample (ibid.: 300), but also if the researchers checked the two design assumptions of the ICT (no-liars and no-design-effect). Additional the different versions of the ICT can be coded to look at whether a double list procedure (Glynn (2013)) performs more often better than a single list ICT.

A completely different potential approach would be to take a deeper look at the items of the ICT with qualitative methods, for instance to see how the non-sensitive items are constructed compared to the sensitive and to code patterns. Or future work can focus on the validation of such indirect questions techniques. While it is already proven that the CM reports false positives (Höglinger and Jann (2018)), the ICT has no such verification checks, therefor future research could focus on validation techniques on the ICT.

Bibliography

Ahart, A. M., & Sackett, P. R. (2004). A new method of examining relationships between individual difference measures and sensitive behavior criteria: Evaluating the unmatched count technique. *Organizational Research Methods*, *7*(1), 101–114.

Ahlquist, J. S. (2018). List experiment design, non-strategic respondent error, and item count technique estimators. *Political Analysis*, *26*(1), 34–53.

Aichholzer, J., Kritzinger, S., Wagner, M., & Zeglovits, E. (2014). How has radical right support transformed established political conflicts? the case of austria. *West European Politics*, *37*(1), 113–137.

Aronow, P. M., Coppock, A., Crawford, F. W., & Green, D. P. (2015). Combining list experiment and direct question estimates of sensitive behavior prevalence. *Journal of survey statistics and methodology*, *3*(1), 43–66.

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2016). *Multivariate analysemethoden*. Springer.

Biemer, P., & Brown, G. (2005). Model-based estimation of drug use prevalence using item count data. *Journal of Official Statistics*, *21*(2), 287.

Blair, G., Coppock, A., & Moor, M. (2018). When to worry about sensitivity bias: evidence from 30 years of list experiments. *Work. Pap., Univ. Calif., Los Angeles, CA Google Scholar Article Location*.

Blair, G., & Imai, K. (2012). Statistical analysis of list experiments. *Political Analysis*, *20*(1), 47–77.

Bogner, K., & Landrock, U. (2015). Antworttendenzen in standardisierten umfragen. *Mannheim, GESIS–Leibniz Institut für Sozialwissenschaften (SDM Survey Guidelines)*.

Boruch, R. F. (1971). Assuring confidentiality of responses in social research: A note on strategies. *The American Sociologist*, *6*, 308–311.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: The definitive guide to questionnaire design–for market research, political polls, and social and health questionnaires*. Hoboken, NJ: John Wiley & Sons.

Canan, C. E. (2017). *Prescription analgesic use and misuse among people living with hiv in the united states* (Unpublished doctoral dissertation). Johns Hopkins University.

Cannell, C. F., Miller, P. V., & Oksenberg, L. (1981). Research on interviewing techniques. *Sociological Methodology*, *12*, 389–437.

Cannell, C. F., Oksenberg, L., & Converse, J. M. (1977). Striving for response accuracy: Experiments in new interviewing techniques. *Journal of Marketing Research*, *14*(3), 306–315.

Çarkoğlu, A., & Aytaç, S. E. (2015). Who gets targeted for vote-buying? evidence from an augmented list experiment in turkey. *European Political Science Review*, *7*(4), 547–566.

Chaudhuri, A. (2016). *Randomized response and indirect questioning techniques in surveys*. London: Chapman & Hall/CRC.

Chaudhuri, A., & Christofides, T. (2013). *Indirect questioning in sample surveys*. London: Springer.

Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the Randomized Response model. *Psychological Methods*, *3*(2), 160.

Cohen, J. (1988). *Statistical power analysis*. Hillsdale, NJ: Erlbaum.

Comşa, M., & Postelnicu, C. (2012). Measuring social desirability effects on self-reported turnout using the item count technique. *International Journal of Public Opinion Research*, *25*(2), 153–172.

Coutts, E., & Jann, B. (2011). Sensitive questions in online surveys: Experimental results for the randomized response technique (rrt) and the unmatched count technique (uct). *Sociological Methods & Research*, *40*(1), 169–193.

Coutts, E., Jann, B., Krumpal, I., & Näher, A.-F. (2011). Plagiarism in student papers: Prevalence estimates using special techniques for sensitive questions. *Jahrbücher für Nationalökonomie und Statistik*, *231*(5-6), 749–760.

Dalton, D. R., Wimbush, J. C., & Daily, C. M. (1994). Using the unmatched count technique (uct) to estimate base rates for sensitive behavior. *Personnel Psychology*, *47*(4), 817–829.

Davis, E. O., Crudge, B., Lim, T., O'Connor, D., Roth, V., Hunt, M., & Glikman, J. A. (2019). Correction: Understanding the prevalence of bear part consumption in cambodia: A comparison of specialised questioning techniques. *PloS one*, *14*(3), e0214392.

De Jonge, C. P. K., & Nickerson, D. W. (2014). Artificial inflation or deflation? assessing the item count technique in comparative surveys. *Political Behavior*, *36*(3), 659–682.

De Leeuw, E. D., Hox, J., & Dillman, D. (2012). *International handbook of survey methodology*. Routledge.

Diekmann, A. (2012). Making use of "Benford's Law" for the Randomized Response Technique. *Sociological Methods & Research*, *41*(2), 325–334.

Droitcour, J., Caspar, R. A., Hubbard, M. L., Parsley, T. L., Visscher, W., & Ezzati, T. M. (1991). The Item Count Technique as a method of indirect questioning: A review of its development and a case study application. In *Measurement errors in surveys* (pp. 185–210). New York, NY: John Wiley & Sons.

Fanelli, D., Costas, R., & Larivière, V. (2015). Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. *PLoS One*, *10*(6), e0127556.

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 665–694.

Flavin, P., & Keane, M. (2009). How angry am i? let me count the ways: Question format bias in list experiments. *Unpublished Manuscript*, 1–17.

Flesch, R. (1948). A new readability yardstick. *Journal of applied psychology*, *32*(3), 221.

Fowler Jr, F. J., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error* (Vol. 18). Sage.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*(6203), 1502–1505.

Frye, T., Gehlbach, S., Marquardt, K. L., & Reuter, O. J. (2017). Is putin's popularity real? *Post-Soviet Affairs*, *33*(1), 1–15.

Gaia, A., & Al Baghal, T. (2019). The longitudinal item count technique: a new technique for asking sensitive questions in surveys. *Methods, data, analyses: a journal for quantitative methods and survey methodology (mda)*, *13*(1), 111–137.

Glynn, A. N. (2013). What can we learn with statistical truth serum? design and analysis of the list experiment. *Public Opinion Quarterly*, *77*(S1), 159–172.

Gonzalez-Ocantos, E., De Jonge, C. K., Meléndez, C., Osorio, J., & Nickerson, D. W.

(2012). Vote buying and social desirability bias: Experimental evidence from nicaragua. *American Journal of Political Science*, *56*(1), 202–217.

Grant, T., Moon, R., & Gleason, S. (2012). *Asking many, many sensitive questions: A person-count method for social desirability bias.* Slides and Paper presented at 37th Annual Conference of the Midwest Association for Public Opinion Research, Chicago, IL.

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question Randomized Response model: Theoretical framework. *Journal of the American Statistical Association*, *64*(326), 520–539.

Gschwend, T., Juhl, S., & Lehrer, R. (2018). Die 'Sonntagsfrage', soziale Erwünschtheit und die AfD: Wie alternative Messmethoden der Politikwissenschaft weiterhelfen können. *Politische Vierteljahresschrift*, 1–27.

Hanmer, M. J., & Ozan Kalkan, K. (2013). Behind the curve: Clarifying the best approach to calculating predicted probabilities and marginal effects from limited dependent variable models. *American Journal of Political Science*, *57*(1), 263–277.

Hesselmann, F. (2018). Science and its others: examining the discourse about scientific misconduct through a postcolonial lens. *Identities*, 1–19.

Hesselmann, F., Graf, V., Schmidt, M., & Reinhart, M. (2017). The visibility of scientific misconduct: A review of the literature on retracted journal articles. *Current sociology*, *65*(6), 814–845.

Hesselmann, F., Wienefoet, V., & Reinhart, M. (2014). Measuring scientific misconduct—lessons from criminology. *Publications*, *2*(3), 61–70.

Hinsley, A., Keane, A., St. John, F. A., Ibbett, H., & Nuno, A. (2019). Asking sensitive questions using the unmatched count technique: Applications and guidelines for conservation. *Methods in Ecology and Evolution*, *10*(3), 308–319.

Hoffmann, A., de Puiseau, B. W., Schmidt, A. F., & Musch, J. (2017). On the comprehensibility and perceived privacy protection of indirect questioning techniques. *Behavior research methods*, *49*(4), 1470–1483.

Hoffmann, A., Diedenhofen, B., Verschuere, B., & Musch, J. (2015). A strong validation of the Crosswise Model using experimentally-induced cheating behavior. *Experimental Psychology*, *62*(6), 403–414.

Hoffmann, A., & Musch, J. (2016). Assessing the validity of two indirect questioning techniques: A stochastic lie detector versus the crosswise model. *Behavior Research Methods*, *48*(3), 1032–1046.

Höglinger, M. (2016). *Revealing the truth? validating the randomized response technique for surveying sensitive topics* (Unpublished doctoral dissertation). ETH Zurich.

Höglinger, M., & Diekmann, A. (2017). Uncovering a blind spot in sensitive question research: False positives undermine the Crosswise-Model RRT. *Political Analysis*, *25*, 131–137.

Höglinger, M., & Jann, B. (2018). More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model. *PloS one*, *13*(8), 1–22.

Höglinger, M., Jann, B., & Diekmann, A. (2014). Sensitive questions in online surveys: An experimental evaluation of the Randomized Response Technique and the Crosswise Model. *University of Bern Social Sciences Working Paper No*, *9*.

Höglinger, M., Jann, B., & Diekmann, A. (2016). Sensitive questions in online surveys: An experimental evaluation of different implementations of the Randomized Response Technique and the Crosswise Model. *Survey Research Methods*, *10*(3), 171–187.

Holbrook, A. L., Green, M. C., & Krosnick, J. A. (2003). Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public Opinion Qarterly*, *67*(1), 79–125.

Holbrook, A. L., & Krosnick, J. A. (2009). Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly*, *74*(1), 37–67.

Holbrook, A. L., & Krosnick, J. A. (2010). Measuring voter turnout by using the randomized response technique: Evidence calling into question the method's validity. *Public Opinion Quarterly*, *74*(2), 328–343.

Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, *30*(2), 161–172.

Hopp, C., & Speil, A. (2018). Estimating the extent of deceitful behaviour using crosswise elicitation models. *Applied Economics Letters*, 1–5.

Horvitz, D. G., Simmons, W. R., & Shah, B. (1968). Unrelated question Randomized Response Model. *Journal of the American Statistical Association*, *63*(322), 754–754.

Houston, J., & Tran, A. (2001). A survey of tax evasion using the Randomized Response Technique. In *Advances in taxation* (pp. 69–94). Bingley: Emerald.

Hubbard, M. L., Casper, R. A., Lessler, J. T., et al. (1989). Respondents' reactions to item count lists and randomized response. *Proceedings of the Survey Research Section', American Statistical Association, Washington, DC*, 544–448.

Imai, K. (2011). Multivariate regression analysis for the item count technique. *Journal of the American Statistical Association*, *106*(494), 407–416.

Jann, B., Jerke, J., & Krumpal, I. (2012). Asking sensitive questions using the Crosswise Model: An experimental survey measuring plagiarism. *Public Opinion Quarterly*, 32–49.

Jerke, J., Johann, D., Rauhut, H., & Thomas, K. (2019, Forthcoming). Too sophisticated even for highly educated survey respondents? a qualitative assessment of indirect question formats for sensitive questions. *Survey Research Methods*.

Johann, D., & Thomas, K. (2017). Testing the validity of the Crosswise Model: A Study on attitudes towards Muslims. *Survey Methods: Insights from the Field*, 10.13094/SMIF-2017-00001.

Johann, D., Thomas, K., Faas, T., & Fietkau, S. (2016). Alternative messverfahren rechtspopulistischen wählens im vergleich: Empirische erkenntnisse aus deutschland und österreich. In *Wahlen und wähler* (pp. 447–470). Springer.

Junkermann, J., Wolter, F., & Ehler, I. (2019). *A comprehensive meta-analysis of experimental survey studies on the performance of the item count technique.* Paper presented at 8th European Survey Research Association Conference, Zagreb, HR.

Khosravi, A., Mousavi, S. A., Chaman, R., Khosravi, F., Amiri, M., & Shamsipour, M. (2015). Crosswise model to assess sensitive issues: a study on prevalence of drug abuse among university students of iran. *International journal of high risk behaviors & addiction*, *4*(2).

Kim, S. H., & Kim, S. (2016). Social desirability bias in measuring public service motivation. *International Public Management Journal*, *19*(3), 293–319.

Korndörfer, M., Krumpal, I., & Schmukle, S. C. (2014). Measuring and explaining tax

evasion: Improving self-reports using the Crosswise Model. *Journal of Economic Psychology*, *45*, 18–32.

Kraemer, H. C., & Blasey, C. (2015). *How many subjects? Statistical power analysis in research*. London: Sage.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213–236.

Krumpal, I. (2012). Estimating the prevalence of Xenophobia and Anti-semitism in Germany: A comparison of Randomized Response and Direct Questioning. *Social Science Research*, *41*(6), 1387–1403.

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity*, *47*(4), 2025–2047.

Krumpal, I., Jann, B., Auspurg, K., & von Hermanni, H. (2015). Asking sensitive questions: A critical account of the Randomized Response Technique and related methods. In *Improving survey methods: Lessons from recent research* (pp. 122–136). New York, NY: Routledge.

Kuhn, P. M., & Vivyan, N. (2018). Reducing turnout misreporting in online surveys. *Public opinion quarterly.*.

Kuk, A. Y. (1990). Asking sensitive questions indirectly. *Biometrika*, *77*(2), 436–438.

Kuklinski, J. H., Cobb, M. D., & Gilens, M. (1997). Racial attitudes and the" new south". *The Journal of Politics*, *59*(2), 323–349.

Kundt, T. C. (2014). *Applying Benford's Law to the Crosswise Model: Findings from an online survey on tax evasion* (Tech. Rep. No. 148). Hamburg: Helmut-Schmidt-University.

Kundt, T. C., Misch, F., & Nerré, B. (2016). Re-assessing the merits of measuring tax evasion through business surveys: An application of the Crosswise Model. *International Tax and Public Finance*, *24*(1), 112—133.

Kundt, T. C., Misch, F., & Nerré, B. (2017). Re-assessing the merits of measuring tax evasion through business surveys: An application of the Crosswise Model. *International Tax and Public Finance*, *24*(1), 112—133.

LaBrie, J. W., & Earleywine, M. (2000). Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique. *Journal of Sex Research*, *37*(4), 321–326.

Lensvelt-Mulders, G. (2008). Surveying sensitive topics. *International handbook of survey methodology*, *46*, 41.

Lensvelt-Mulders, G., Hox, J. J., Van der Heijden, P. G., & Maas, C. J. (2005a). Meta-analysis of Randomized Response research: Thirty-five years of validation. *Sociological Methods & Research*, *33*(3), 319–348.

Lensvelt-Mulders, G., Hox, J. J., Van der Heijden, P. G., & Maas, C. J. (2005b). Meta-analysis of randomized response research thirty-five years of validation. *Sociological Methods & Research*, *33*(3), 319–348.

Lenzner, T., Kaczmirek, L., & Lenzner, A. (2010). Cognitive burden of survey questions and response times: A psycholinguistic experiment. *Applied cognitive psychology*, *24*(7), 1003–1020.

Lenzner, T., & Menold, N. (2015). Frageformulierung (version 1.1). *Mannheim: GESIS–Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines)*.

Li, Y. (2019). Relaxing the no liars assumption in list experiment analyses. *Political Analysis*, 1–16.

Lippitt, M., Reese Masterson, A., Sierra, A., Davis, A. B., & White, M. A. (2014). An exploration of social desirability bias in measurement of attitudes toward breastfeeding in public. *Journal of Human Lactation*, *30*(3), 358–366.

Mangat, N. S. (1994). An improved Randomized Response strategy. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*, 93–95.

Marlowe, D., & Crowne, D. P. (1964). *The approval motive: Studies in evaluative dependence*. Wiley New York.

Martinez, M. D., & Craig, S. C. (2010). Race and 2008 presidential politics in florida: A list experiment. In *The forum* (Vol. 8).

Martinson, B. C., Anderson, M. S., & De Vries, R. (2005). Scientists behaving badly. *Nature*, *435*(7043), 737.

Miles, J., & Shevlin, M. (2001). *Applying regression and correlation: A guide for students and researchers*. London: Sage.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior*. George Washington University.

Moseson, H., Massaquoi, M., Dehlendorf, C., Bawo, L., Dahn, B., Zolia, Y., ... Gerdts, C. (2015). Reducing under-reporting of stigmatized health events using the list

experiment: results from a randomized, population-based study of abortion in liberia. *International journal of epidemiology*, *44*(6), 1951–1958.

Moshagen, M., Hilbig, B. E., Erdfelder, E., & Moritz, A. (2014). An experimental validation method for questioning techniques that assess sensitive issues. *Experimental Psychology*, *61*(1), 48.

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests*. Hove: Routledge.

Nakhaee, M. R., Pakravan, F., & Nakhaee, N. (2013). Prevalence of use of anabolic steroids by bodybuilders using three methods in a city of Iran. *Addiction & Health*, *5*(3-4), 77–81.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology*, *15*(3), 263–280.

Nepusz, T., Petróczi, A., Naughton, D. P., Epton, T., & Norman, P. (2014). Estimating the prevalence of socially sensitive behaviors: Attributing guilty and innocent noncompliance with the single sample count method. *Psychological methods*, *19*(3), 334.

Nuno, A., Bunnefeld, N., Naiman, L. C., & MILNER-GULLAND, E. J. (2013). A novel approach to assessing the prevalence and drivers of illegal bushmeat hunting in the serengeti. *Conservation Biology*, *27*(6), 1355–1365.

Nuno, A., & John, F. A. S. (2015). How to ask sensitive questions in conservation: A review of specialized questioning techniques. *Biological Conservation*, *189*, 5–15.

Ostapczuk, M., Musch, J., & Moshagen, M. (2009). A Randomized-Response investigation of the education effect in attitudes towards foreigners. *European Journal of Social Psychology*, *39*(6), 920–931.

Patrzek, J., Sattler, S., van Veen, F., Grunschel, C., & Fries, S. (2015). Investigating the effect of academic procrastination on the frequency and variety of academic misconduct: a panel study. *Studies in Higher Education*, *40*(6), 1014–1029.

Petróczi, A., Nepusz, T., Cross, P., Taft, H., Shah, S., Deshmukh, N., ... others (2011). New non-randomised model to assess the prevalence of discriminating behaviour: a pilot study on mephedrone. *Substance abuse treatment, prevention, and policy*, *6*(1), 20.

Phillips, D. L., & Clancy, K. J. (1972). Some effects of "Social Desirability" in survey

studies. *American Journal of Sociology*, *77*(5), 921–940.

Porst, R. (2009). Question wording–zur formulierung von fragebogen-fragen. In *Fragebogen* (pp. 95–114). Springer.

Preisendörfer, P., & Wolter, F. (2014). Who is telling the truth? a validation study on determinants of response behavior in surveys. *Public Opinion Quarterly*, *78*(1), 126–146.

Raghavarao, D., & Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(1), 40–45.

Rayburn, N. R., Earleywine, M., & Davison, G. C. (2003). An investigation of base rates of anti-gay hate crimes using the unmatched-count technique. *Journal of Aggression, Maltreatment & Trauma*, *6*(2), 137–152.

Redlawsk, D. P., Tolbert, C. J., & Franko, W. (2010). Voters, emotions, and race in 2008: Obama as the first black president. *Political Research Quarterly*, *63*(4), 875–889.

Roberts, D. L., & John, F. A. S. (2014). Estimating the prevalence of researcher misconduct: a study of uk academics within biological sciences. *PeerJ*, *2*, e562.

Rosenfeld, B., Imai, K., & Shapiro, J. N. (2016). An empirical validation study of popular survey methodologies for sensitive questions. *American Journal of Political Science*, *60*(3), 783–802.

Safiri, S., Rahimi-Movaghar, A., Mansournia, M. A., Yunesian, M., Shamsipour, M., Sadeghi-Bazargani, H., & Fotouhi, A. (2019). Sensitivity of crosswise model to simplistic selection of nonsensitive questions: an application to estimate substance use, alcohol consumption and extramarital sex among iranian college students. *Substance use & misuse*, *54*(4), 601–611.

Schwarz, N., Knäuper, B., Oyserman, D., & Stich, C. (2008). The psychology of asking questions. *International handbook of survey methodology*, 18–22.

Seibert, L. J. (2019). Conducting member surveys with sensitive topics. *Association Metrics*.

Shamsipour, M., Yunesian, M., Fotouhi, A., Jann, B., Rahimi-Movaghar, A., Asghari, F., & Akhlaghi, A. A. (2014). Estimating the prevalence of illicit drug use among students using the Crosswise Model. *Substance Use & Misuse*, *49*(10),

1303–1310.

Sheppard, S. C., & Earleywine, M. (2013). Using the unmatched count technique to improve base rate estimates of risky driving behaviours among veterans of the wars in iraq and afghanistan. *Injury prevention*, *19*(6), 382–386.

Smith, N. F., & Street, D. J. (2003). The use of balanced incomplete block designs in designing randomized response surveys. *Australian & New Zealand Journal of Statistics*, *45*(2), 181–194.

Tan, M. T., Tian, G.-L., & Tang, M.-L. (2009). Sample surveys with sensitive questions: A nonrandomized response approach. *The American Statistician*, *63*, 9–16.

Thomas, K., Johann, D., Kritzinger, S., Plescia, C., & Zeglovits, E. (2017). Estimating sensitive behavior: The ICT and high-incidence electoral behavior. *International Journal of Public Opinion Research*, *29*(1), 157–171.

Thomas, K., Schnell, J., R., & Noack, M. (2019). *Education and income effects in randomised response estimates of undeclared employment.* Paper presented at 8th European Survey Research Association Conference, Zagreb, HR.

Tourangeau, R. (1984). Cognitive sciences and survey methods. In T. J. Jabine (Ed.), *Cognitive aspects of survey methodology: Building a bridge between disciplines* (pp. 73–100). Washington D.C.: National Academy Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response.* Cambridge University Press.

Tourangeau, R., & Smith, T. W. (1996). Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public opinion quarterly*, *60*(2), 275–304.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin*, *133*(5), 859–883.

Trappmann, M., Krumpal, I., Kirchner, A., & Jann, B. (2014). Item Sum: A new technique for asking quantitative sensitive questions. *Journal of Survey Statistics and Methodology*, *2*(1), 58–77.

Tsuchiya, T. (2005). Domain estimators for the item count technique. *Survey Methodology*, *31*(1), 41–51.

Tsuchiya, T., Hirai, Y., & Ono, S. (2007). A study of the properties of the item count technique. *Public Opinion Quarterly*, *71*(2), 253–272.

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of Randomized Response models. *Psychological methods*, *17*(4), 623–641.

Umesh, U. N., & Peterson, R. A. (1991). A critical evaluation of the Randomized Response method applications, validation, and research agenda. *Sociological Methods & Research*, *20*(1), 104–138.

Vakilian, K., Keramat, A., Mousavi, S. A., & Chaman, R. (2019). Experience assessment of tobacco smoking, alcohol drinking, and substance use among shahroud university students by crosswise model estimation–the alarm to families. *The Open Public Health Journal*, *12*(1).

Vakilian, K., Mousavi, S. A., & Keramat, A. (2014). Estimation of sexual behavior in the 18-to-24-years-old Iranian youth based on a Crosswise Model study. *BMC Research Notes*, *7*(1), 1–4.

Vakilian, K., Mousavi, S. A., Keramat, A., & Chaman, R. (2016). Knowledge, attitude, self-efficacy and estimation of frequency of condom use among Iranian students based on a Crosswise Model. *International Journal of Adolescent Medicine and Health*(10.1515/ijamh-2016-0010), 1–5.

Walzenbach, S., & Hinz, T. (2019). Pouring water into wine: Revisiting the advantages of the crosswise model for asking sensitive questions. *Survey Methods: Insights from the Field*, 16.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*(309), 63–69.

Waubert de Puiseau, B., Hoffmann, A., & Musch, J. (2017). How indirect questioning techniques may promote democracy: A preelection polling experiment. *Basic and Applied Social Psychology*, *39*(4), 209–217.

Wolter, F., & Laier, B. (2014a). The effectiveness of the Item Count Technique in eliciting valid answers to sensitive questions. An evaluation in the context of self-reported delinquency. In *Survey research methods* (Vol. 8, pp. 153–168).

Wolter, F., & Laier, B. (2014b). The effectiveness of the item count technique in eliciting valid answers to sensitive questions. An evaluation in the context of self-reported delinquency. In *Survey research methods* (Vol. 8, pp. 153–168).

Wolter, F., & Preisendörfer, P. (2013). Asking sensitive questions an evaluation of

the Randomized Response Technique versus direct questioning using individual validation data. *Sociological Methods & Research*, *42*(3), 321–353.

Yan, T., & Cantor, D. (2019). Asking survey questions about criminal justice involvement. *Public Health Reports*, *134*(1_suppl), 46S–56S.

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, *67*(3), 251–263.

Zeglovits, E., & Kritzinger, S. (2013). New attempts to reduce overreporting of voter turnout and their effects. *International Journal of Public Opinion Research*, *26*(2), 224–234.

Zimmerman, R. S., & Langer, L. M. (1995). Improving estimates of prevalence rates of sensitive behaviors: The randomized lists technique and consideration of self-reported honesty. *Journal of Sex Research*, *32*(2), 107–117.

A Additional thoughts, variables and analysis

Including all studies in the full analysis

First idea was to take every study with a questionnaire into account in the analysis, independent of their original language. For this I contacted multiple researchers and asked for permission or a copy to the original questionnaire, so the instruction can be extracted from there. After multiple follow-ups only 26 responded positively. This proceedings turned out to be impossible with respect to the deadline of this thesis, so only the original English studies are included in the analysis.

The first mail draft was the following: `Dear xy,`

`I am currently working on my Master Thesis, a meta-analysis of studies which either use the Item Count Technique (which is also known as the Unmatched Count Technique, the List Experiment or a Survey Experiment) or the Crosswise Model to shed some light on how different operationalisations affect the performance of these techniques.`

`With great interest, I have read your paper about 'xx'. I would like to include it as a part of my analysis.`

`In order to answer my research questions I would ideally need the questionnaire you used in your study. If that's not possible then I am particularly interested in the precise wording of the instruction you used to introduce the special technique and all sensitive and nonsensitive items or questions, that have been used in the study. The english translations of the instruction and of the items are sufficient.`

`I would be very thankful for your cooperation. Please don't hesitate to reach out to me if you have any questions.`

`Sincerely,`
`Antonia Velicu`

`-`

`antonia.velicu@uzh.ch`

`Degree program: Master of Arts in Social Science (Sociology)`

`Institute of Sociology`

`University of Zurich`

Additional Variables and Analysis

The number of sentences, number of comma, uniqueness of the words, number of syllables, ratio syllables pro words are created and in the end excluded variables. The codes extract in different ways multiple factors which were not theoretically relatable, so I decided to leave it and focus on a few but central variables to use in the regression model. For further supplementary material accompanying this analysis, please contact author: antonia.velicu@uzh.ch The number of sentences, number of words and number of syllables go in the same direction of trying to capture the length of an instruction, number of words is a bit more accurate. Number of commas makes no sense or no significant meaning in the English language. Uniqueness of words alone is just a ratio build on the nlp-package, how unique a single word is. I decided against this because capturing the repetition of words makes more sense - and they are co linear. Ratio syllables per words has not a huge variance, as its language specific. this variable would make more sense while comparing different languages, but in the English language alone its a redundant and therefore dischargeable information.

The Flesch-Kincaid readability test by Flesch (1948) is the United States Military Standard for assessing the difficulty of technical manuals. It measures the readability on a scale of the appropriate school level.

$$x = 206.835 - 1.015(\frac{totalwords}{totalsentences}) - 84.6(\frac{totalsyllables}{totalwords})$$

I calculated the test for every instruction, but in the end decided against its inclusion, because it appears, that the rest is not applicable on short text passages.

I also coded the non-sensitive topic of the Crosswise model, but as the variance was not huge, I excluded it.

I also tried to do a factor analysis with the instruction variables. The results were inconclusive and in the end I excluded the two factors again.

B Further descriptive analysis

Crosswise Model

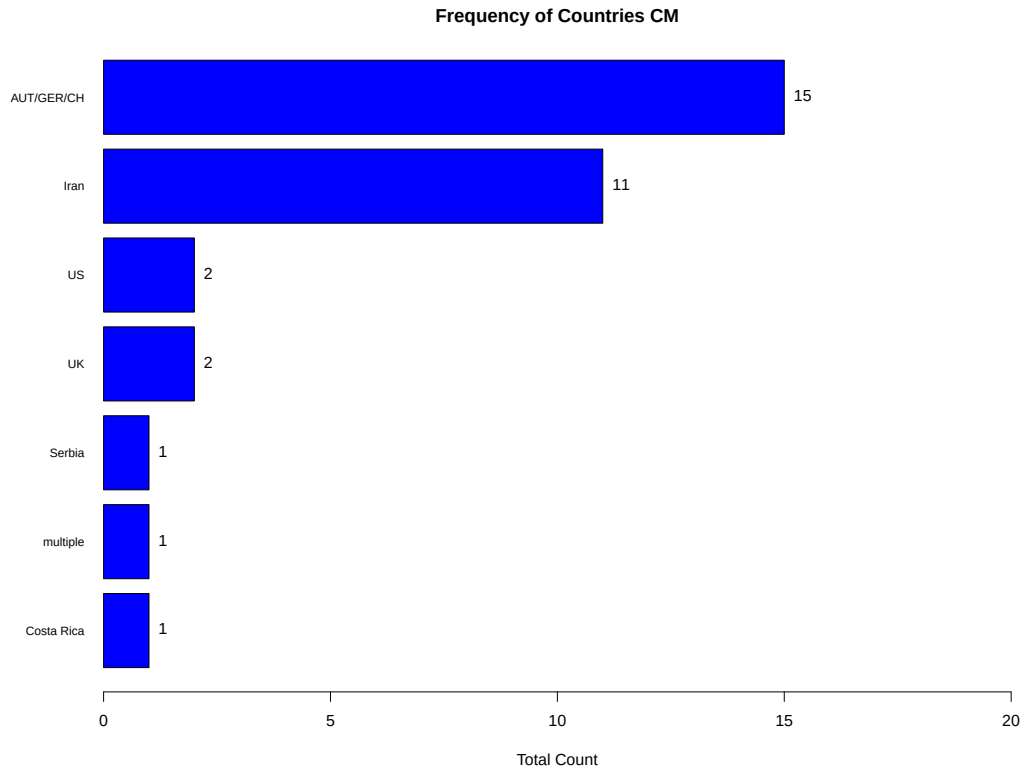


Figure 12: Variety of Countries (CM)

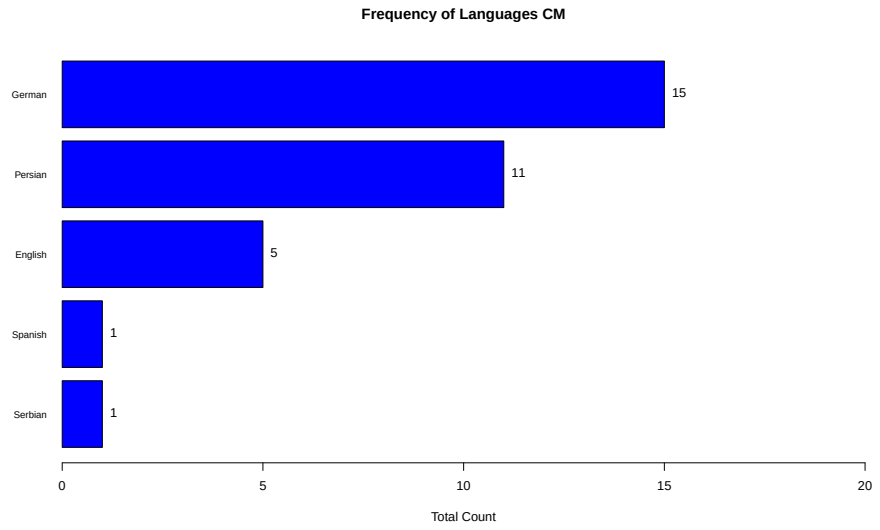


Figure 13: Variety of Languages (CM)

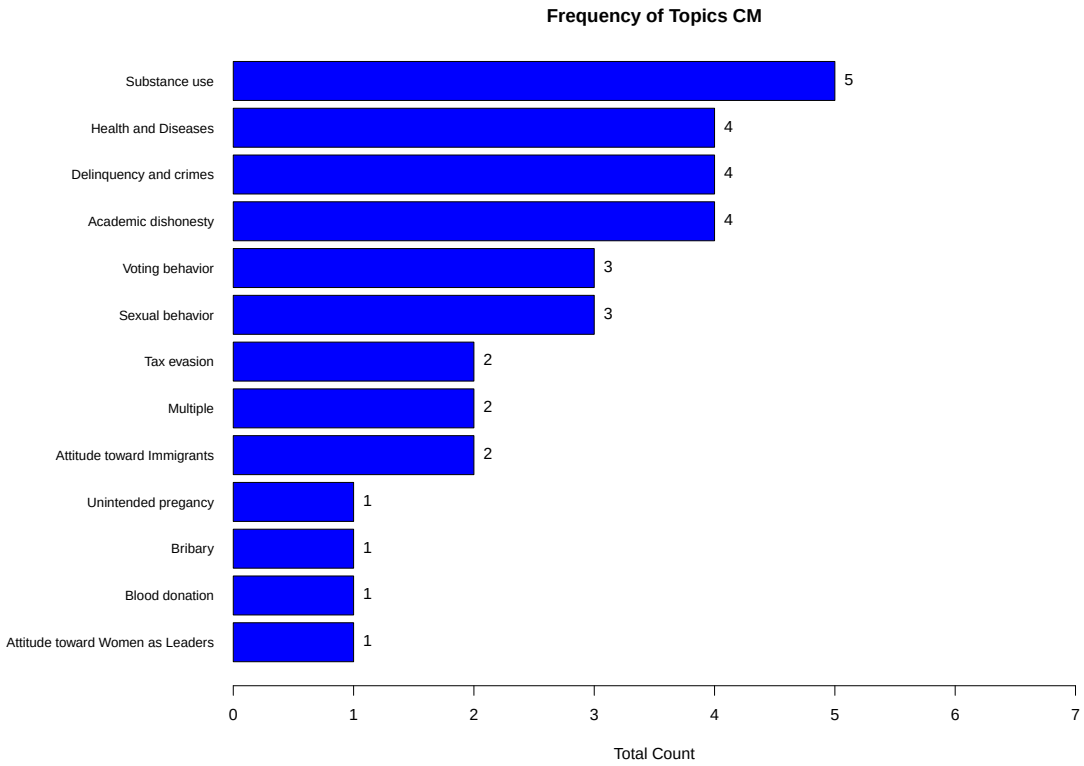


Figure 14: Variety of Topics (CM)

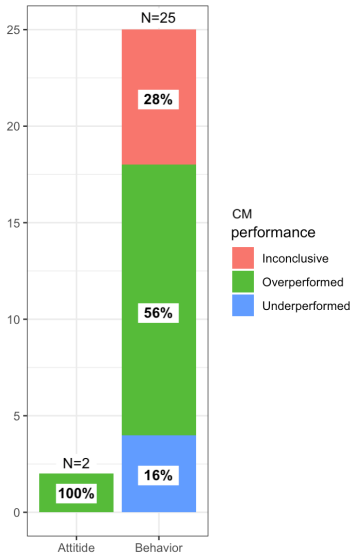


Figure 15: Asking a sensitive behavior is semi-successful.

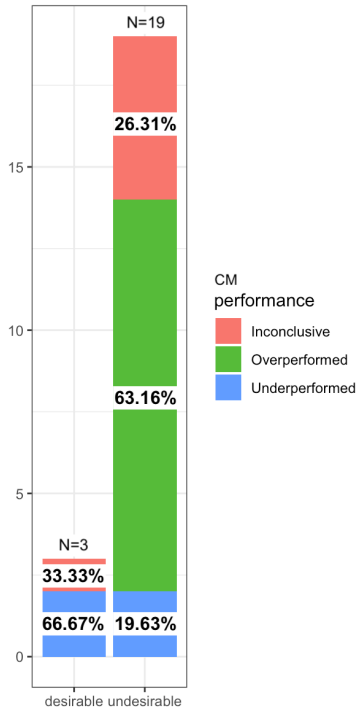


Figure 16: In CM studies the comparison is nonsensical, undesirable items appear to perform in three out of five cases better.

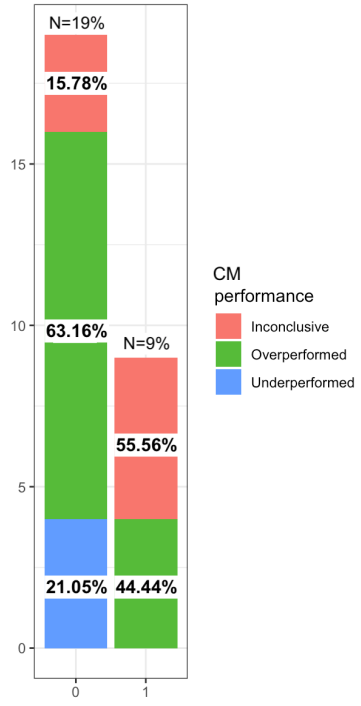


Figure 17: Surprisingly non pretested CM studies appear to work better.

Item Count Technique

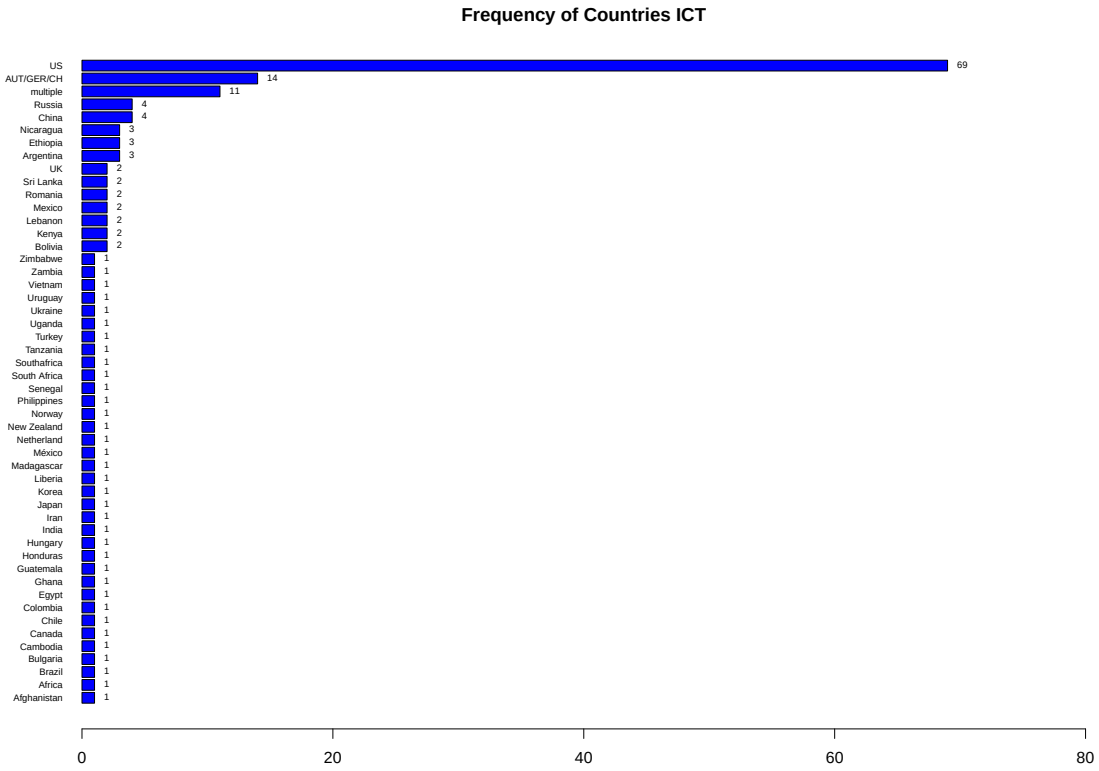


Figure 18: Variety of Countries (ICT)

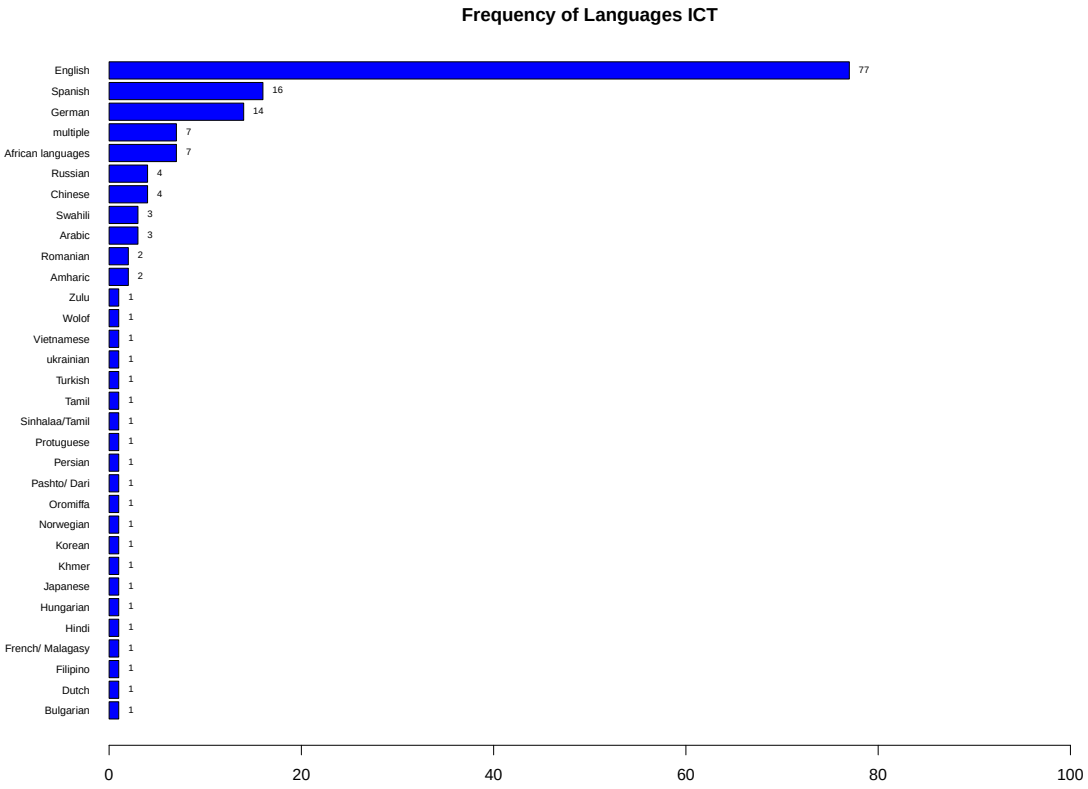


Figure 19: Variety of Languages (ICT)

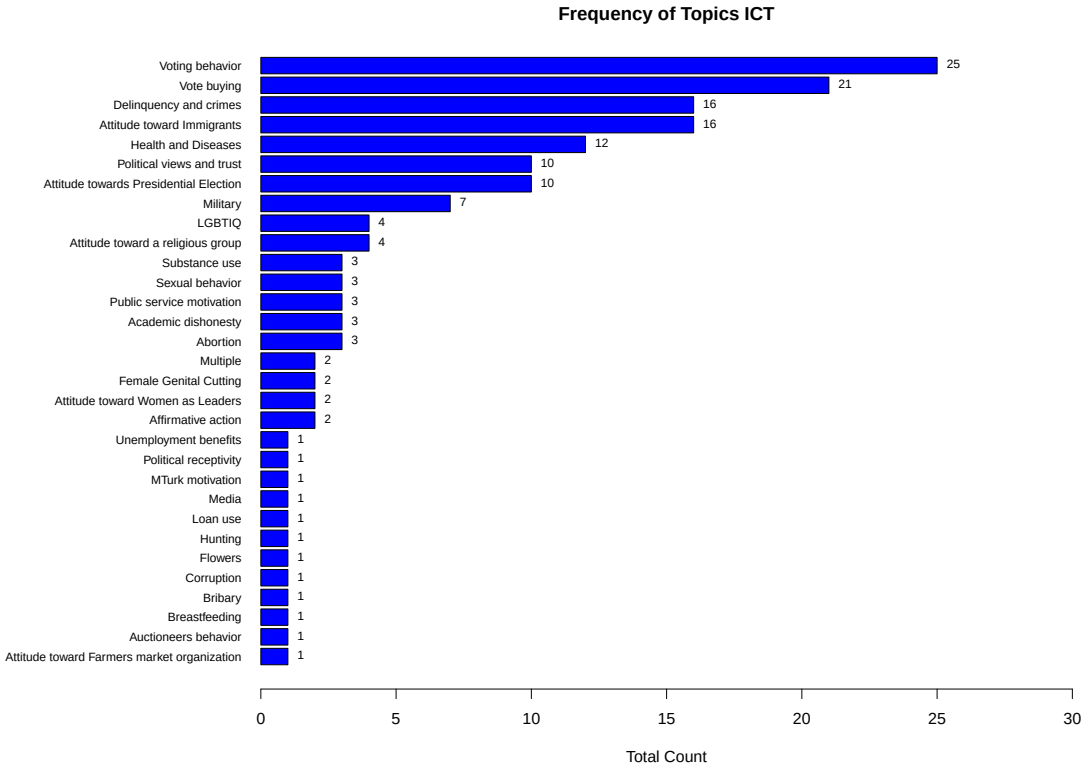


Figure 20: Variety of Topics (ICT)

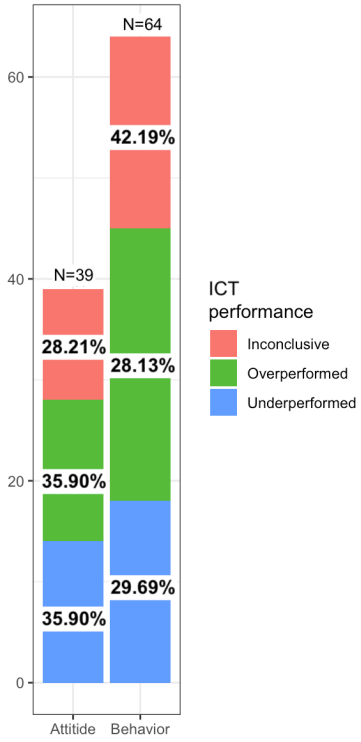


Figure 21: Asking a delicate attitude through the ICT appears to be the same as asking a precarious behavior, but there is almost twice the amount of observations in the behavior category.

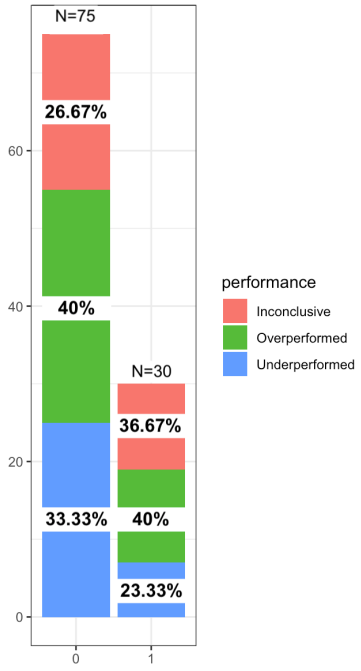
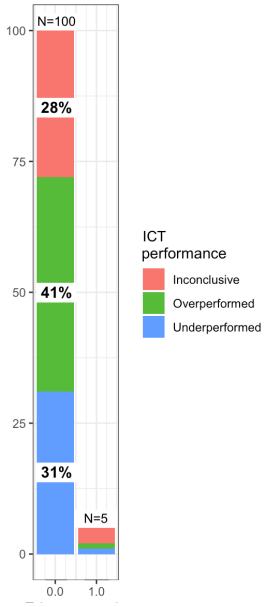
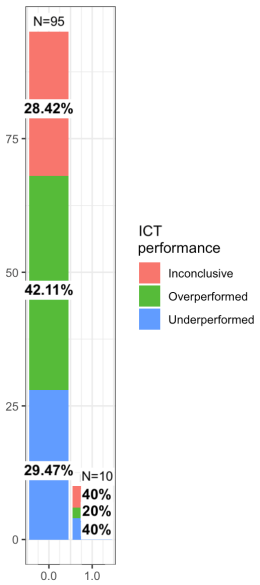


Figure 22: Conducting or refraining from a pretest appears to have only hardly any effect on the outcome of the comparison between ICT and DQ.



(a) Statistical Explanation



(b) Use of an Example

Figure 23: In the cases of ICT studies, a comparison can not to be drawn as the sample sizes differ significantly.

C Further analysis

Bivariate analysis

To look at it from a different angle, I also tried to analyse the statistical explanation and use of an exmaple together (see Table 3). Out of 160 ICT studies, only four (2.5%) use both, an example and a in depth explanation and 146 (91.25%) have refrained from using any of the two. The CM studies show similar results, of the 33 studies in this sample, four (12.12%) have decided to go with both, an example to illustrate the method and a statistical explanation of how the method works and 18 (54.55%) avoid either of the above. At the first glance at the analysis shows that not many studies used more tools to explain the techniques, or didn't report it. This could be an indication that researchers do not worry too much about the unconventional structure of the techniques.

ICT	$\mathbf{1}^{\mathbf{expl}}$	$\mathbf{0}^{\mathbf{expl}}$	$\mathbf{sum}^{\mathbf{exam}}$	**CM**	$\mathbf{1}^{\mathbf{expl}}$	$\mathbf{0}^{\mathbf{expl}}$	$\mathbf{sum}^{\mathbf{exam}}$
$\mathbf{1}^{\mathbf{exam}}$	4	9	*13*	$\mathbf{1}^{\mathbf{exam}}$	4	3	*7*
$\mathbf{0}^{\mathbf{exam}}$	1	146	*137*	$\mathbf{0}^{\mathbf{exam}}$	8	18	*26*
$\mathbf{sum}^{\mathbf{expl}}$	*5*	*155*	*160*	$\mathbf{sum}^{\mathbf{expl}}$	*12*	*21*	*33*

Table 3: Statistical explanation and use of an example: Not many ICT and CM used an example or an explanation for a better comprehension of the unconventional technique.

Multivariate analysis

Variables	Model 1***			Model 2		
	Coeff	*Odds Ratio*	*p*	*Coeff*	*Odds Ratio*	*p*
Constant	2.27 (1.37)	9.66 (13.26)	0.098	2.25 (1.37)	9.45 (12.93)	0.101
Democratic country	-1.48 (1.28)	0.28 (.029)	0.247	-1.46 (1.27)	.023 (0.296)	0.251
Socially undesirable	1.77** (0.89)	5.86 (5.22)	0.047	1.77** (0.89)	5.85** (5.20)	0.047
Attitude	-0.62 (0.84)	0.54 (0.46)	0.456	-0.64 (0.84)	0.52 (0.44)	0.446
Survey Mode (Ref.: F-t-f)						
Self administered	-3.69*** (1.56)	0.02 (0.04)	0.018	-3.55*** (1.59)	0.03*** (0.05)	0.026
Telephone	-0.70 (1.37)	0.49 (0.68)	0.608	-0.7 (1.37)	0.48 (0.66)	0.595
Online	-2.60*** (1.20)	0.07 (0.09)	0.030	-2.58*** (1.20)	0.08*** (0.09)	0.032
Pretest	-1.13 (.90)	0.322 (0.33)	0.211	-1.08 (0.91)	0.34 (0.31)	0.236
Quantity non-s. I. (Ref.: 4)						
3 non-sens. Items	-0.52 (.98)	0.59 (0.59)	0.598	-0.53 (0.98)	0.59 (0.58)	0.589
>5 non-sens. Items	-1.34 (1.25)	0.26 (0.33)	0.282	-1.39 (1.29)	0.248 (0.32)	0.279
Context Items	-0.51 (.80)	0.599 (0.48)	0.524	-0.44 (0.82)	0.65 (0.53)	0.593
Explanation				-	-	-
Example				-1.22 (2.01)	0.296 (0.62)	0.558
N	*65*			*65*		
McFadden's Pseudo R2	*0.361*			*0.356*		
Count R2	*0.800*			*0.800*		
Log likelihood	*-28.189*			*-27.999*		

Table 4: Results of Logistic Regression on Method Level

Variables	Model 4		
	Coeff	*Odds Ratio*	*p*
Constant	57.27 (1584.242)	7.42e+24 (1.18e+28))	0.971
Democratic country	-	-	-
Socially undesirable	5.24* (3.09)	189.51 (586.11)	0.090
Attitude	3.38 (2.55)	29.34 (74.75)	0.185
Survey Mode (Ref.: F-t-f)			
Self administered	-21.62 (1584.11)	4.09e-10 (3.11e-07)	0.989
Telephone	-6.44 (4.17)	0.002 (0.007)	0.122
Online	-22.351 (1584.11)	01.96e-10 (3.11e-07)	0.989
Pretest	-2.92 (2.91)	0.05 (0.157)	0.316
Quantity non-s. I. (Ref.: 4)			
3 non-sens. Items	-1.10 (1.71)	0.33 (0.57)	0.518
>5 non-sens. Items	-.079 (1.32)	0.92 (1.22)	0.952
Context Items	5.09* (3.01)	162.49 (488.46)	0.090
Explanation	-	-	-
Example	-	-	-
Words/sentences	-0.50 (0.35)	0.60 (0.209)	0.145
Word repetitions	-32.90** (16.39)	5.13e-15 (8.42e-14)	0.045
Number of Words	-0.24* (0.13)	0.788 (0.105)	0.076
N	*54*		
McFadden's Pseudo R2	*0.115*		
Count R2	*0.661*		
Log likelihood	*-36.020*		

Table 5: Results of Logistic Regression on Item Level

hlineheightVariables	M 1.1***	M 1.2	M 1.3	M 1.4***	M 1.5	M 1.6***	M 1.7***	Model 1***	M 2.1	M 2.2	Model 2***	M 3.1	Model 3
Constant	1.03***	.70	-.32	.69	-.38	-.39	-1.01***	2.31	-.35	-.30	2.23	1.56	-.73
	(.52)	(.64)	(.25)	(.39)	(.24)	(.31)	(.30)	(1.37)*	(.20)*	(.21)	(1.37)	(5.37)	(15.90)
Democratic country	-1.70**							-1.47			-1.45		-
	(.57)							(1.27)			(1.26)		
Socially undesirable		-0.92						1.70			1.74		-
		(.59)						(.90)*			(.90)*		
Attitude			-.22					-.53			-.63		-4.32
			(.42)					(.85)			(.84)		(3.82)
Survey Mode (Ref.: F-t-f)													
Self administered				-2.49***				-3.70***			-3.51		.59
				(.86)				(1.56)			(1.60)**		(2.34)
Telephone				-.69				-.73			-.72		1.46
				(.69)				(1.38)			(1.37)		(2.65)
Online				-1.89***				-2.68***			-2.55 (1.21)**		-
				(.53)				(1.21)					
Pretest					-.023			-1.18			-1.07		-1.89
					(.44)			(.90)			(.91)		(2.42)
Quantity non-s. I. (Ref.: 4)													
3 non-sens. Items						.44		-.54			-.51		1.99
						(.45)		(.996)			(.98)		(2.56)
>5 non-sens. Items						.-1.35*		-1.61			-1.39		.059
						(.70)		(1.23)			(1.28)		(2.19)
Context Items							1.29***	-.48			-.42		.027
							(.43)	(.80)			(.82)		(2.14)
Statistical Explanation									-1.04		-		-
									(1.14)				
Example										-1.1	-1.22		-
										(.82)	(2.07)		
Wording Instruction													
Ratio words per sentence												-.071	.32
												(.12)	(.35)
Word repetitions												-1.23	-4.38
												(4.98)	(13.77)
Number words												-.02	-.022
												(.038)	(.14)
N	*99*	*79*	*102*	*99*	*104*	*99*	*98*	*68*	*104*	*104*	*69*	*38*	*18*
McFadden's Pseudo R2	*0.0753*	*0.0118*	*0.0021*	*0.1433*	*0.0000*	*0.0594*	*0.0706*	*0.3730*	*0.0071*	*0.0148*	*0.3583*	*0.0190*	*0.1981*
Count R2	*0.677*	*0.582*	*0.598*	*0.596*	*0.616*	*0.663*	*0.663*	*0.812*	*0.596*	*0.596*	*0.800*	*0.737*	*0.778*
Log likelihood	*-61.263*	*-51.845*	*-67.466*	*-56.047*	*-69.092*	*-62.089*	*-60.588*	*-28.394*	*-68.866*	*-68.318*	*-27.871*	*-21.085*	*-8.504*

Table 6: Results of Logistic Regression - Models with each a complete sample on method-level

Variables	M 4.1*	M 4.2	M 4.3	M 4.4**	M 4.5	M 4.6	M 4.7	M 4.8	Model 4*
Constant	-1.39	.049	-.36	.42	-.37	.08	.25	7.81	57.27
	(1.12)	(.31)	(1.28))	(.28)	(.43)	(.29)	(.25)	(5.46)	(1584.24)
Democratic country	-	-	-	-	-	-	-	-	-
Social desirable	1.72								5.24*
	(1.15)								(3.09)
Attitude		.24							3.38
		(.49)							(2.55)
Survey Mode (Ref.: F-t-f)									
Self administered			.96						-21.62
			(1.32)						(1584.11)
Telephone			-1.15						-6.44
			(.97)						(4.17)
Online			.36						-22.35
			(1.32)						((1584.11)
Pretest				-1.21					-2.92
				(.61)**					(2.91)
Quan. non- s. I. (Ref.: 4))									
3 non- s. Items					.74				-1.10
					(.61)				(1.70)
>5 non- s. Items					.77				.08
					(.59)				(1.32)
Context Items						.18			5.01
						(.51)			(3.01)*
Complexity									
Statistical Explanation	-	-	-	-	-	-	-	-	-
Example							-1.64		-
							(1.15)		
Number words								-.075*	-32.90*
								(.038)	(16.34)
Ratio words/ sentences								-.014	-.50
								(0.65)	(.34)
Word repetitions								-6.25	-.24**
								(4.87)	(.13)
N	*60*	*69*	*68*	*69*	*69*	*69*	*69*	*69*	*54*
Mc Fadden's Pseudo R2	*0.034*	*0.0025*	*0.0267*	*0.0445*	*0.0221*	*0.0012*	*0.0269*	*0.0516*	*0.283*

Table 7: Results of Logistic Regression with all Variables on Item Level

D List of studies

Paper	Method	Country	Topic
Ahart/ Sackett 2004 A New Method of Examining Relationships Between Individual Difference Measures and Sensitive Behavior Criteria: Evaluating the Unmatched Count Technique	ICT	US	Multiple
Ahlquist/ Mayer/ Jackman 2014 Alien Abduction and Voter Impersonation in the 2012 Us General Election: Evidence from a Survey List Experiment	ICT	US	Voting behavior
Alvarez 2019 Paying Attention to Inattentive Survey Respondents	ICT	US	
An 2015 The role of social desirability bias and racial/ethnic composition on the relation between education and attitude toward immigration restrictionism	ICT	US	Immigration
Anderson/ Simmons/ Milnes/ Earleywine 2007 Effect of response format on endorsement of eating disordered attitudes and behaviors	ICT	US	Health
Antin/ Shaw 2012 Social Desirability Bias and Self-Reports of Motivation: A Study of Amazon Mechanical Turk in the Us and India	ICT	multiple	MTurk motivation
Arentoft et al 2016 Comparing the unmatched count technique and direct self-report for sensitive health-risk behaviors in HIV+ adults	ICT	US	Health
Aronow/ Coppock/ Crawford/ Freen 2015 Combining list experiment and direct question estimates of sensitive behavior prevalence	ICT	multiple	
Ash 2013 Identity Group Allegiance in Civil Wars.	ICT	Lebanon	Military
Banayejeddi 2019 Implementation evaluation of an iron supplementation programme in high-school students: the crosswise model	CM	Iran	
Bauer 2019 A Nudge in a New Direction: Integrating Behavioral Economic Strategies Into Suicide Prevention Work	ICT	US	Health and Diseases
Becerra Mizuno 2012 Does Everyone Have a Price? The Demand Side of Clientelism and Vote-Buying in an Emerging Democracy	ICT	México	Vote buying
Benson/ Merolla/ Geer 2011 Two Steps Forward, One Step Back? Bias in the 2008 Presidential Election	ICT	US	Presidential election
Berinsky 2018 Telling the Truth About Believing the Lies? Evidence for the Limited Prevalence of Expressive Survey Responding	ICT	US	Politics
Biemer/ Brown 2005 Model-based estimation of drug use prevalence using item count data	ICT	US	Substance use
Blair/ Imai/ Lyall 2014 Comparing and Combining List and Endorsement Experiments: Evidence from Afghanistan	ICT	Afghanistan	Military
Bøttkjær 2017 Crying Wolf: An Experimental Test of the Augmented List Experiment	ICT	multiple	Vote buying
Bratton/ Dulani/ Masunungure 2016 Detecting Manipulation in Authoritarian Elections: Survey-Based Methods in Zimbabwe	ICT	Zimbabwe	Voting behavior
Brierley 2017 Politicians and Bureaucrats: The Politics of Development and Corruption in Ghana	ICT	Ghana	Corruption
Brooke 2017 Sectarianism and Social Conformity: Evidence from Egypt	ICT	Egypt	Racism
Brown-Iannuzzi 2019 The Illusion of Political Tolerance: Social Desirability and Self- Reported Voting Preferences	ICT	US	
Brownback/ Novotny 2018 Social Desirability Bias and Polling Errors in the 2016 Presidential Election	ICT	US	Politics
Burden/ Ono/ Yamada 2017 Reassessing Public Support for a Female President	ICT	US	Presidential election
Canan 2017 Prescription analgesic use and misuse among people living with HIV in the US	CM	US	Health
Cappelen/ Mitbo 2016 Intra-Eu Labour Migration and Support for the Norwegian Welfare State	ICT	Norway	Immigration

Carkoglu/ Aytac 2015 Who gets targeted for vote-buying? Evidence from an augmented list experiment in Turkey	ICT	Turkey	Vote buying
Coffman / Coffman/ Ericson 2017 The Size of the LGBT Population and the Magnitude of Antigay Sentiment Are Substantially Underestimated	ICT	multiple	LGBTIQ
Comsa /Postelnicu 2013 Measuring Social Desirability Effects on Self-Reported Turnout Using the Item Count Technique	ICT	Romania	Voting behavior
Conley/ McCabe 2011 Body Mass Index and Physical Attractiveness: Evidence from a Combination Image- Alteration/List Experiment	ICT	US	Health
Copoock 2017 Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-Representative List Experiment	ICT	US	Presidential election
Corbacho/ Gingerich/ Oliveros / Ruiz-Vega 2016 Corruption as a Self-Fulfilling Prophecy: Evidence from a Survey Experiment in Costa Rica	CM	Costa Rica	Bribary
Costange 2017 Clientelism in Competitive and Uncompetitive Elections	ICT	Lebanon	Vote buying
Coutts/Jann 2011 Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT)	ICT	AUT/GER/CH	Multiple
Coutts/Jann/Krumpal/Näher 2011 Plagiarism in student papers: Prevalence estimates using special techniques for sensitive questions	Both	AUT/GER/CH	Academic dishonesty
Cowan/ Wu/ Makela/ England 2016 Alternative Estimates of Lifetime Prevalence Of Abortion from Indirect Survey Questioning Methods	ICT	US	Abortion
Creighton 2018 Race, Wealth and the Masking of Oppositionto Immigrants in the Netherlands	ICT	Netherland	
Creighton/ Jamal / Malancu 2015 Has Opposition to Immigration Increased in the United States after the Economic Crisis? An Experimental Approach	ICT	US	Racism
Creighton/Jamal 2015 Does Islam play a role in anti-immigrant sentiment? An experimental approach	ICT	US	Immigration
Cruz 2014 Buying One Vote at a Time or Buying in Bulk? Politician Networks and Electoral Strategies	ICT	Philippines	Vote buying
Dalton/ Wimbush/ Daily 1994 Using the unmatched count technique (UCT) to estimate base rates for sensitive behavior	ICT	US	
David 2019 Understanding the prevalence of bear part consumption in Cambodia: A comparison of specialised questioning techniques	ICT	Cambodia	
De Jonge 2015 Who Lies About Electoral Gifts?	ICT	multiple	Vote buying
DeCao / Lutz 2018 Sensitive Survey Questions: Measuring Attitudes Regarding Female Genital Cutting Through a List Experiment	ICT	Ethiopia	Female Genital Cutting
Droitcour et al. 1991 The Item Count Technique as a method of indirect questioning: A review of its development and a case study application	ICT	US	Health
Druckman/ Gilli/ Klar/ Robinson 2015 Measuring Drug and Alcohol Use Among College Student-Athletes	ICT	US	Substance use
Eady 2017 The Statistical Analysis of Misreporting on Sensitive Survey Questions	ICT	Canada	Voting behavior
Enzmann 2017 Die Anwendbarkeit des Crosswise-Modells zur Prüfung kultureller Unterschiede sozial erwünschten Antwortverhaltens: Implikationen für seinen Einsatz in internationalen Studien zu selbstberichteter Delinquenz.	CM	US	Delinquency and crimes
Eriksen/ Lutz/ Tadesse 2018 Social Desirability, Opportunism and Actual Support for Farmers' Market Organisations in Ethiopia	ICT	Ethiopia	Farmers market organization
Eslami et al 2013 Importance of Pre-pregnancy Counseling in Iran: Results from the High Risk Pregnancy Survey 2012	CM	Iran	Abortion
Flavin/ Keane 2009 How Angry am I? Let Me Count the Ways: Question Format Bias in List Experiment	ICT	US	Presidential election

Frye 2019 Hitting Them With Carrots: Voter Intimidation and Vote Buying in Russia	ICT	Russia	
Frye/ Gehlbach/ Marquardt/ Reuter 2017 Is Putin's popularity real?	ICT	Russia	Presidential election
Frye/ Reuter/ Szakonyi 2017 Political Machines at Work Voter Mobilization and Electoral Subversion in the Workplace	ICT	Russia	Voting buying
Gibson 2018 Indirect questioning method reveals hidden support for female genital cutting in South Central Ethiopia	ICT	Ethiopia	Female Genital Cutting
Gilens/ Sniderman/ Kuklinski 1998 Affirmative Action and the Politics of Realignment	ICT	US	Affirmative action
Gimpel/ Hui 2016 Inadvertent and Intentional Partisan Residential Sorting	ICT	US	
Gonzalez-Ocantos et al 2012 Vote Buying and Social Desirability Bias:Experimental Evidence from Nicaragua	ICT	Nicaragua	Voting behavior
Gosen 2014 Social Desirability in Survey Research: Can the List Experiment Provide the Truth?	ICT	AUT/GER/CH	
Gunarathne et al 2016 Is Hiding Foot and Mouth Disease Sensitive Behavior for Farmers? A Survey Study in Sri Lanka	ICT	Sri Lanka	Health
Haber et al 2018 List randomization for eliciting HIV status and sexual behaviors in rural KwaZulu-Natal, South Africa: a randomized experiment using known true values for validation	ICT	Southafrica	Health
Hadji et al. 2016 Assessing the Prevalence of Publication Misconduct Among Iranian Authors Using a Double List Experiment	ICT	Iran	Academic dishonesty
Harden 2013 Multidimensional Responsiveness: The Determinants of Legislators' Representational Priorities	ICT	US	
Harisson 2015 Profiling unauthorized natural resource users for better targeting of conservation interventions	ICT	Uganda	Multiple
Harris 2018 The Economic Roots of Anti-Immigrant Prejudice in the Global South: Evidence from South Africa	ICT	South Africa	Voting behavior
Heerwig/ McCabe 2009 Education and Social Desirability Bias: The Case of a Black Presidential Candidate	ICT	US	Presidential election
Hinsley/ Nuno/ Ridout/ John/ Roberts 2016 Estimating the Extent of CITES Noncompliance among Traders and End-Consumers; Lessons from the Global Orchid Trade	ICT	US	Flowers
Hoffman/ Musch 2016 Assessing the validity of two indirect questioning techniques: A Stochastic Lie Detector versus the Crosswise Model	CM	AUT/GER/CH	Racism
Hoffmann/ Musch 2018 Prejudice against women leaders: Insights from an indirect questioning approach	CM	AUT/GER/CH	Gender
Hoffmann/ Diedenhofen/ Verschuere/ Musch 2015 A Strong Validation of the Crosswise Model Using Experimentally-Induced Cheating Behavior	CM	AUT/GER/CH	
Höglinger 2017 Uncovering a blind spot in sensitive question research: False positives undermine the Crosswise-Model RRT	CM	AUT/GER/CH	Health
Höglinger/ Jann 2018 More is not always better: An experimental individual-level validation of the randomized response technique and the crosswise model	CM	multiple	Multiple
Höglinger/ Jann/ Diekmann 2016 Sensitive Questions in Online Surveys: An Experimental Evaluation of Different Implementations of the Randomized Response Technique and the Crosswise Model	CM	AUT/GER/CH	Academic dishonesty
Holbrook/ Krosnick 2009 Social desirability bias in voter turnout reports: Tests using the item count technique	ICT	US	Voting behavior
Hopp/Speil 2018 Estimating the extent of deceitful behaviour using crosswise elicitation models	Both	AUT/GER/CH	Multiple
Imai/ Park/ Greene 2015 Using the Predicted Responses from List Experiments as Explanatory Variables in Regression Models	ICT	Mexico	Vote buying

Jann/Jerke/Krumpal 2012 Asking sensitive questions using the Crosswise Model: An experimental survey measuring plagiarism	CM	AUT/GER/CH	Academic dishonesty
Janus 2010 The Influence of Social Desirability Pressures on Expressed Immigration Attitudes	ICT	US	Racism
Johann/Thomas 2017 Testing the Validity of the Crosswise Model: A Study on Attitudes Towards Muslims	CM	AUT/GER/CH	Racism
Johann/Thomas/ Faas/ Fietkau 2016 Alternative Messverfahren rechtspopulistischen Wählens im Vergleich:Empirische Erkenntnisse aus Deutschland und Österreich	ICT	AUT/GER/CH	Voting behavior
Kalinin 2016 The social desirability bias in autocrat's electoral ratings: evidence from the 2012 Russian presidential elections	ICT	Russia	Voting behavior
Kane/ Craig/ Wald 2004 Religion and Presidential Politics in Florida: A List Experiment	ICT	US	Presidential election
Karlan/ Zinman 2010 List randomization for sensitive behavior: An application for measuring use of loan proceeds	ICT	multiple	Loan use
Khosravi et al 2015 Crosswise Model to Assess sensitive issues - a study on prevalence of drug abuse among university students of iran	CM	Iran	Substance use
Kim/ Kim 2016 Social Desirability Bias in Measuring Public Service Motivation	ICT	Korea	Public service motivation
Kim/ Kim 2016 National Culture and Social Desirability Bias in Measuring Public Service Motivation	ICT	multiple	Public service motivation
Kim/ Kim 2017 Ethnic Differences in Social Desirability Bias: Effects on the Analysis of Public Service Motivation	ICT	US	Public service motivation
Kirchner/ Krumpal/ Trappmann/ von Hermanni 2013 Messung und Erklärung von Schwarzarbeit in Deutschland – Eine empirische Befragungsstudie unter besonderer Berücksichtigung des Problems der sozialen Erwünschtheit	ICT	AUT/GER/CH	Delinquency and crimes
Kleykamp/ Hipes/ MacLean 2018 Who Supports Us Veterans and Who Exaggerates Their Support?	ICT	US	Military
Klimas 2019 Higher testosterone levels are associated with unfaithful behavior in men	CM	AUT/GER/CH	Sexual behavior
Knoll 2013 Assessing the Effect of Social Desirability on Nativism Attitude Responses	ICT	US	Racism
Körndorfer/Krumpal/Schmukle 2014 Measuring and explaining tax evasion: Improving self-reports using the Crosswise Model	CM	AUT/GER/CH	Tax evasion
Kramon 2016 Where Is Vote Buying Effective? Evidence from a List Experiment in Kenya	ICT	Kenya	Vote buying
Kramon 2019 (Mis)Measuring Sensitive Attitudes with the List Experiment: Solutions to List Experiment Breakdown in Kenya	ICT	Kenya	
Krebs/ Linquist/ Warner/ Fosher/ Martin/ Childers 2011 Comparing Sexual Assault Prevalence Estimates Obtained With Direct and Indirect Questioning Techniques	ICT	US	Delinquency and crimes
Kuha/ Jackson 2014 The item count method for sensitive survey questions: modelling criminal behaviour	ICT	multiple	Delinquency and crimes
Kuhn / Vivyan 2018 Reducing Turnout Misreporting in Online Surveys	Both	UK	Voting behavior
Kuklinski et al. 1997 Racial prejudice and attitudes toward afrmative action	ICT	US	Racism
Kuklinski/ Cobb/ Gliens 1997 Racial Attitudes and the "New South"	ICT	US	Racism
Kundt/Misch/Nerré 2013 Re-assessing the merits of measuring tax evasion through business surveys: An application of the Crosswise Model	CM	Serbia	Tax evasion
LaBrie/ Earleywine 2000 Sexual risk behaviors and alcohol: Higher base rates revealed using the unmatched-count technique	ICT	US	Sexual behavior

Lavender/Anderson 2007 Effect of response format on endorsement of eating disordered attitudes and behaviors	ICT	US	Health
Lax/ Phillips/ Stollwerk 2016 Are Survey Respondents Lying about Their Support for Same-Sex Marriage? Lessons from a List Experiment	ICT	US	LGBTIQ
Lehrer/ Juhl/ Gschwend 2019 The wisdom of crowds design for sensitive survey questions	Both	AUT/GER/CH	Voting behavior
Li/ Shi/ Zhu 2018 The Face of Internet Recruitment: Evaluating the Labor Markets of Online Crowdsourcing Platforms in China	ICT	China	Politics
Lippitt /Masterson /Sierra /Davis /White 2014 An Exploration of Social Desirability Bias in Measurement of Attitudes toward Breastfeeding in Public	ICT	US	Breastfeeding
Malesky/ Gueorgulev/ Jensen 2015 Monopoly Money: Foreign Investment and Bribery in Vietnam, a Survey Experiment	ICT	Vietnam	Bribary
Mares/ Muntean/ Petrova 2018 Economic Intimidation in Contemporary Elections: Evidence from Romania and Bulgaria	ICT	multiple	Voting behavior
Mares/ Muntean/ Petrova 2017 Pressure, Favours, and Vote-Buying: Experimental Evidence from Romania and Bulgaria	ICT	multiple	Vote buying
Mares/ Young 2018 The Core Voter's Curse: Coercion and Clientelism in Hungarian Elections	ICT	Hungary	Vote buying
Martinez/ Craig 2010 Race and 2008 Presidential Politics in Florida: A List Experiment	ICT	US	Presidential election
Matanock/ Garcia-Sanchez 2017 Does Counterinsurgent Success Match Social Support? Evidence from a Survey Experiment in Colombia	ICT	Colombia	Military
McKenzie/ Siegel 2013 Eliciting illegal migration rates through list randomization	ICT	multiple	Delinquency and crimes
Meng/ Pan/ Yang 2017 Conditional Receptivity to Citizen Participation: Evidence From a Survey Experiment in China	ICT	China	Politics
Mirzazadeh et al 2018 Underreporting in HIV-Related High-Risk Behaviors: Comparing the Results of Multiple Data Collection Methods in a Behavioral Survey of Prisoners in Iran	CM	Iran	Multiple
Moseson et al 2015 Reducing under-reporting of stigmatized health events using the List Experiment: results from a randomized, population-based study of abortion in Liberia	ICT	Liberia	Abortion
Moseson/ Gerdts 2017 Measuring Texas women's experiences with abortion self-induction using a list experiment	ICT	US	Abortion
Muralidharan/ Niehaus/ Sukhatankar 2016 Building State Capacity: Evidence from Biometric Smartcards in India	ICT	India	Unemployment benefits
Nakhaee/ Pakravan/ Nakhaee 2013 Prevalence of use of anabolic steroids by bodybuilders using three methods in a city of Iran	CM	Iran	Substance use
Nasirian et al 2018 Does Crosswise Method Cause Overestimation? An Example to Estimate the Frequency of Symptoms Associated With Sexually Transmitted Infections in General Population: A Cross Sectional Study	CM	Iran	Health
Nuno et al 2013 A novel approach to assessing the prevalence and drivers of illegal bushmeat hunting in the Serengeti	ICT	Tanzania	Hunting
Oliveros 2016 Making It Personal: Clientelism, Favors, and the Personalization of Public Administration in Argentina	ICT	Argentina	Politics
Pavao 2015 The Failures of Electoral Accountability for Corruption: Brazil and Beyond. University of Notre Dame	ICT	Brazil	Delinquency and crimes
Pechenkino/Bausch/Skinner 2018 The Pitfalls of List Experiments in Conflict Zones	ICT	Ukraine	Military
Peterman/ Palermo/ Handa/ Seidenfeld 2017 List randomization for soliciting experience of intimate partner violence: Application to the evaluation of Zambia's unconditional child grant program	ICT	Zambia	

Prior 2009 Improving Media Effects Research Through Better Measurement of News Exposure	ICT	US	Media
Randrianantorandro/ Kono/ Kubota 2015 Knowledge and behavior in an animal disease outbreak - Evidence from the item count technique in a case of African swine fever in Madagascar	ICT	Madagascar	Health
Rayburn 2003 An investigation of base rates of anti-gay hate crimes using the unmatched-count technique	ICT	US	LGBTIQ
Rayburn/ Earleywine /Davison 2003 Base Rates of Hate Crime Victimization among College Students	ICT	US	Delinquency and crimes
Redlawsk/ Tolbert/ Franko 2010 Voters, Emotions, and Race in 2008: Obama as the First Black President	ICT	US	Presidential election
Roberts/John 2014 Estimating the prevalence of researcher misconduct: a study of UK academics within biological sciences	Both	UK	Academic dishonesty
Robinson 2019 Self-censorship of regime support in authoritarian states: Evidence from list experiments in China	ICT	China	
Ronconi/ Zarazaga 2015 Labor Exclusion and the Erosion of Citizenship Responsibilities	ICT	Argentina	
Rosenfeld/Imai/Shapiro 2016 An Empirical Validation Study of Popular Survey Methodologies for Sensitive Questions	ICT	US	Voting behavior
Safiri et al. 2018 Sensitivity of Crosswise Model to Simplistic Selection of Non-sensitive Questions: An Application to Estimate Substance Use, Alcohol Consumption and Extramarital Sex Among Iranian College Students	CM	Iran	Multiple
Schnapp 2019 Sensitive Question Techniques and Careless Responding: Adjusting the Crosswise Model for Random Answers	CM	AUT/GER/CH	Health
Seljan/ Lochner/ Gold/ Davis 2016 I Know What You Did Last Cycle: Improving the Detection of State Campaign Finance Violations.	ICT	US	Politics
Shamsipour et al 2014 Estimating the prevalence of illicit drug use among students using the Crosswise Model	CM	Iran	Substance use
Sheppard /Earleywine 2013 Using the unmatched count technique to improve base rate estimates of risky driving behaviours among veterans of the wars in Iraq and Afghanistan	ICT	US	Military
Starosta/ Earleywine 2014 Assessing Base Rates of Sexual Behavior Using the Unmatched Count Technique	ICT	US	Sexual behavior
Streb 2008 Social desirability effects and support for a female American president	ICT	US	Presidential election
Su 2015 Improving Health Measures: Evidence from a List Experiment, Cognitive Interviews, and a Vignette Study	ICT	China	Health
Swedlund 2017 Can Foreign Aid Donors Credibly Threaten to Suspend Aid? Evidence from a Cross-National Survey of Donor Officials	ICT	Africa	Health
Thomas/ Gavin / Milfont 2015 Estimating non-compliance among recreational fishers: Insights into factors affecting the usefulness of the randomized response and item count techniques	ICT	NZ	Delinquency and crimes
Thomas/ Johann / Kritzinger / Plescia / Zeglovits 2016 Estimating sensitive behavior: The ICT and high-incidence electoral behavior	ICT	AUT/GER/CH	Voting behavior
Traunmüller 2019 The Silent Victims of Sexual Violence during War: Evidence from a List Experiment in Sri Lanka	ICT	Sri Lanka	
Treibich/ Lepine Estimating 2019 misreporting in condom use and its determinants among sex workers: Evidence from the list randomisation method	ICT	Senegal	Sexual behavior
Tsuchiya/ Hirai/ Uno 2007 A study of the properties of the item count technique	ICT	Japan	Multiple
Vakilian/ Kermat/ Mousavi/ Chaman 2019 Experience Assessment of Tobacco Smoking, Alcohol Drinking, and Substance Use Among Shahroud University Students by Crosswise Model Estimation –The Alarm to Families	CM	Iran	Substance use

Vakilian/ Mousavi / Keramat / Chaman 2016 Knowledge, attitude, self-efficacy and estimation of frequency of condom use among Iranian students based on a Crosswise Mode	CM	Iran	Sexual behavior
Vakilian/Mousavi/Kermat 2014 Estimation of sexual behavior in the 18-to-24-years-old Iranian youth based on a Crosswise Model study	CM	Iran	Sexual behavior
Viernich 2018 What's left unsaid? In-group solidarity and ethnic and racial differences in opposition to immigration in the United States	ICT	US	Immigration
Walsh/ Braithwaite 2008 Self-Reported Alcohol Consumption and Sexual Behavior in Males and Females: Using the Unmatched-CountTechnique to Examine Reporting Practices of Socially Sensitive Subjects in a Sample of University Students.	ICT	US	Substance use
Walzenbach 2019 Pouring water into wine: Revisiting the advantages of the crosswise model for asking sensitive questions	CM	AUT/GER/CH	
Waubert de Puiseau/Hoffmann / Musch 2017 How Indirect Questioning Techniques May Promote Democracy: A Preelection Polling Experiment	CM	AUT/GER/CH	Voting behavior
Wilde 2014 Neutral Competence? Polygraphy and Technology-Mediated Administrative Decisions	ICT	US	Racism
Wimbush/Dalton 1997 Base rate for employee theft: Convergence of multiple methods	ICT	US	Delinquency and crimes
Wolter/ Laier 2014 The effectiveness of the Item Count Technique in eliciting valid answers to sensitive questions. An evaluation in the context of self-reported delinquency	ICT	AUT/GER/CH	Delinquency and crimes
Zimmerman/ Langer 1995 Improving estimates of prevalence rates of sensitive behaviors: The randomized lists technique and consideration of self[U+2010]reported honesty	ICT	US	Health